Chapter 2
Science: The Process of Understanding the Natural World and Its Possibilities

What is science? According to the famous Austrian scientist Ernst Mach science is a description of facts of nature that is as complete as possible and as economical (meaning without unnecessary additions) as possible. This definition does not include any word about creativity and does not mention any of the important tools of science. A very important tool in science arises from the notion of similarities and the forming of analogies. Remember the smaller and smaller grains of sand on the beach and the question whether the waves can grind the grains ever smaller. Democrit of Greece thought that there was a limit to the smallness, a smallest type of grain: the atom. He thus had a "theory" about the matter that surrounds us; he believed that everything was made of atoms. Mach did not mention the words theory or logic nor did he mention mathematics in his definition of science. He even said that theories were unnecessary additions (he compared them to dry leaves), and he did refuse to believe in atoms!

Albert Einstein, on the other hand, thought that Mach's definition of science was a bit "stale," and he asked Mach whether he would not find theories useful, if they would lead to the most economical description of nature. Einstein was also pointing toward the fight of Mach with Boltzmann about the existence of atoms, and he clearly sided with Boltzmann and believed that atoms did indeed exist. Since Boltzmann's times, the concept of atoms turned out to be very important and has guided scientist to very successful findings.

If such great luminaries as Boltzmann, Mach, and Einstein had a discussion of what science really is, then we naturally need to be careful when we wish to define it. Therefore, I describe here what science really is, only as far as I understand it myself. Before I do so, I like to say a word about the phrase "as far as I understand it."

I attended an excellent high school in Baden, Austria, and I had a great chemistry teacher. Her name was Marianne Schwarz. She had received a good education at the University of Vienna, and she continued her education by reading textbooks like "The Nature of the Chemical Bond" by Linus Pauling. When she tried to explain how electrons are forming the bonds between the atoms of a molecule, she showed some nice pictures form Pauling's book and said "As far as I understand it, this

K. Hess, *Working Knowledge: STEM Essentials for the 21st Century*,
DOI 10.1007/978-1-4614-3275-3_2, © Springer Science+Business Media New York 2013

means the following." She really impressed me by saying that. Teachers, particularly teachers in Austria at that time, almost never admitted that they did not understand everything fully or could not explain everything to perfection. The reason was and is, of course, that students would use such an admission of ignorance immediately as an excuse that they did not need to know anything about the subject. If a teacher did not really understand it, how and why was a student supposed to know about it? This is a very interesting and important point. All good teachers should, in my opinion, bring the discussion of an important topic toward a satisfactory conclusion. Elementary algebra is, for example, completely understood and can be explained by logical deduction once the axioms are given. Other topics, such as the buzzing of electrons around the nuclei of atoms, or the existence of quarks, cannot be explained with such a degree of certainty and logic, and teachers should admit to the students if a topic cannot yet be completely explained to them. Anything that can not be satisfactorily explained and taught in high school is, of course, not really understood and needs further work by scholars. This is important for the interested students to know, because they are the ones who may find a better explanation later during their best and most productive years. It is also important for any student to know that teachers do not walk on water but try hard to explain what is really known. Students, in turn, need to be impressed if a teacher levels with them and tells them that a subject is not fully understood. They need to be excited that there are always new things to be discovered, and they need to pay particular attention to what the teacher says, instead of flushing things after hearing: "As far as I understand it." So please note that when I say "as far as I understand it," I may not really understand it, and maybe nobody has a totally accepted explanation, but I try my best to explain the current status of understanding.

As far as I understand it, science at its best is what Euclid did. He started with things from everyday life that were useful to measure the size of objects of the surrounding world. If you take a string of a certain length, then you can measure the distance from one place to another. You do this by repeatedly using the string and then you find, for example, that the distance to the neighboring house is hundred times the length of the string. If the length of the string is one "meter," then the distance to the neighboring house is 100 m. Of course, you need first to agree on what one meter is. This is, in principle, a definition. As mentioned, the actual meter measure is made out of platinum and stored in Paris at constant temperature so that it does not expand or contract when the temperature changes. Once one has a measure of distance, one can measure a lot of things. As we know we can then also determine the area of the property that you own. Furthermore, as we know, Euclid worked also with circles, with the length of an arc, and with angles. This initial use of strings and straight objects such as rulers, as well as circles and arcs, was followed by ideas. Euclid used the abstract idea of a point, of a straight line, of an infinitely extended line, of an ideal circle, and so forth. Using these ideas he wrote down simple rules that apply to these ideas, his axioms. Then these simple rules were used to derive theorems, a logical truth, such as the Pythagorean theorem, and the rules together with the theorems permitted us to derive and *predict* a lot of important consequences. We could even calculate the distance from a tower at the

beach to the horizon of the ocean. It is important to note that Euclid used only few (five) axioms and few logical–mathematical rules to derive the laws of geometry. Thus he used a very "restricted" form of language with very careful definitions and much higher precision than we are used to have in ordinary language. This precision is very important for science, and therefore the mathematical–logical language is very important for science.

Of course, the most important point of any scientific approach is to check whether the theoretical results agree with all the observations, with all the actual measurements. Consider our example of calculating the distance to the horizon as seen from a tower (Sect. 1.3.3). To check this calculation one needs to place markers at certain distances out in the ocean and then see whether these markers become visible at the horizon when moving up to certain heights of a tower at the seashore. Of course, this is not an easy experiment. An example of a more straightforward experiment or measurement related to geometrical science would be to measure the area of a triangle by inserting as many little squares into the triangle as possible and by counting the squares and thus measuring the total area. Then we can check if that area is also obtained by the law of Euclid's geometry: multiply the length of the baseline by the height of the triangle and divide by 2. Euclid's geometry was checked in this way over and over in millions and millions of experiments in the thousands of years after Euclid. All these checks came out correct. But no science is ever totally correct, no matter how self-explanatory, no matter how beautiful, logical, and mathematically justified it may be.

More than 2,000 years after Euclid, mathematicians and scientists were still puzzling over Euclid's 5th axiom which says that the sum of the three angles of a triangle equals two right angles. They did not understand whether this should be called an axiom or whether it actually followed from the four other axioms and what it would mean if it were not an axiom or if it were not true? As it turned out, and as we know now from the work of Einstein, the geometry of the universe is not Euclidean and, if very large distances are involved such as those between stars, the sum of the three angles of a triangle does not have to be equal to two right angles. For us on earth, however, it is true to many digits and can be measured to many digits by using laser light to represent the straight lines that form the sides of the triangle. However, we can today also measure, and have measured, how light bends when going around the sun and how then the definition of a straight line becomes more difficult, and how then the 5th axiom can be violated. Details are given in Sect. 5.

To summarize, Euclid's science involved a process of using elements of the world that surrounds us, such as strings and sticks, and then forming limiting abstractions of these elements such as a straight line. Then, using these abstractions, Euclid defined rules relating them to each other. These rules or axioms form the basis of a theoretical framework that can be dealt with using logic and mathematics (e.g., arithmetic and algebra). The results are then carefully checked by measurements with instruments that correspond to the abstractions; for example, a straight line can

be simulated by a laser beam. If a problem is found, and the results of the theory do not agree with experiment, then the theory is corrected and a new improved theory emerges such as Einstein's non-Euclidean theory of the universe.

Euclid's work is often seen as pure mathematics and not really as science. This would be only true if one just takes Euclid's axioms *without their connection to real things* and deduces logical consequences. Indeed, there exists an enormous body of work that has been performed that way and is therefore regarded as pure mathematics. The reason why Euclid's work is so special is that it can be seen as both: as great science connected to all that surrounds us and, on the other hand, as pure logic and mathematics. One can look at Newton's work from a similar point of view: Newton developed the mechanics of planet motion and the corresponding laws of physical science. In the course of this work he discovered and developed the "calculus" which is pure and beautiful mathematics. Science and mathematics are intertwined and gain from each other. Newton's work, however, brings out one more important point of science, a point that was not quite as "visible" in the work of Euclid. This point is that science is *predictive*. The laws that Newton found let us predict the orbit of the planets and where they will be visible on the sky. They also let us predict the path of comets, even comets that we have not seen previously. This predictive quality is the main feature of science, is the feature that makes science useful to mankind.

Many scientific approaches to nature use logic plus some framework of symbols and rules (a theory) and differ quite significantly from the work of Euclid and Newton because they do not use mathematics. The use of mathematics, and a logical framework based on a few axioms, is the signature of scientific maturity and guaranties that precise predictions can be made. The characteristic feature of science is always that the results of the theory, its predictions, can be and are compared to measurements. These measurements connect the theory to nature and prove or disprove the truth content of the theory. If a discrepancy is found then the theory is abandoned or extended until the theory agrees again with the known data that are obtained by observing nature. The observations can be done with our eyes and ears or with elaborate equipment, such as a microscope or telescope, that extends the capabilities of our senses. Thus a theory of science is dealing with abstractions, but clearly connects to nature because its results and predictions have been (and can be) tested over and over by experiments. After many confirming tests, we usually believe that the theory is correct, and then we do not doubt the theory. Indeed, we extend doubt to all that are against the theory. Often this way of thinking is justified, and it is silly to doubt the theorem of Pythagoras when calculating the height of a tower. However, we always need to remember the story of the 5th axiom of Euclid. Einstein found a problem with Euclidean geometry after it had been checked out for more than two thousand years. We also know that Galileo was right, when he refused to believe the then "known fact" that the sun was orbiting around the earth and the earth was standing still; and all his colleagues and even the Pope who opposed him were wrong.

This brings me to the difficult problem of the relation of science and religion to each other. As we can see from Galileo's case, this question can only be

approached with great caution from all sides. We have to give science what belongs to science and to religion what belongs to religion. Under most circumstances this is easy, because science and religion can be clearly separated. Science deals with occurrences in nature that can be experimentally explored and repeated, and science permits us to *predict outcomes* of experiments with a large measure of certainty through the knowledge of some basic laws. If we drop a stone, for example, we can predict with great certainty how fast the stone will fall, and it would be illogical to assume that god will have to govern the falling of the stone by his direct intervention and actions, whenever we choose to drop it. Similarly, if we mix two chemicals and obtain a third one and we can repeat this experiment over and over, then we are exploring natural law and not the directly induced action of an all powerful being. These laws of nature that we explore with scientific methods can be very beautiful and certainly humble us, because our understanding of them is always "anthropomorphic," meaning limited by the human ways of thinking. Boltzmann, the great theoretical physicist, looked at the laws of electromagnetism that were discovered by Faraday and Maxwell and cited the famous verse of Goethe's drama Faust: "Was it a god who wrote these signs?" We can, of course, not prove or disprove the existence of god with our scientific methods. Inversely, the religions of the world cannot deny the existence of a natural law that is accessible to the methods of science and lets us freely experiment and *predict* experimental outcomes.

There are some scientific areas that border on the realm of religion. Scientists have proposed that the universe was created by a "Big Bang" out of an extremely small nucleus and thus virtually out of nothing. Indeed many of the astronomical observations of the universe are consistent with such a theory. There exists, for example, a microwave background radiation in space that could be the remnant of this explosive expansion of the universe. Naturally, religions may assume that the "Big Bang" was the act of creation by god and may discuss this in their instructions. Such themes may also be topics in philosophy, and we admire the wisdom and modesty of Socrates when he said "I know that I know nothing." Science education, however, must exclusively deal with repeatable observations of nature, with deductions of natural laws, and with their justification by further experiments and observations. Naturally, it is important to emphasize that the Big Bang theory is much less convincingly proven than, for example, the laws of falling stones. Scientists cannot now, and never will be able to, perform the crucial experiment, the recreation of the Big Bang. All our evidence can only be indirect, and therefore we are at a borderline between what can be theorized and what can be proven by science.

Such considerations also apply to the often discussed topic of evolution. The theory of evolution was formulated by Charles Darwin and maintains that biological life-forms like birds and humans, and also smaller biological entities, such as biomolecules or cells, have naturally evolved from more humble beginnings to the currently existing forms. Darwin started his book "*On the origin of species*" by explaining how plants were domesticated and made useful by human selection. The natural law, basic to the theory of evolution, is assumed to be the selection of nature and the corresponding survival of the "fittest" or best-suited forms of

life. This selection of nature works similarly to the selection process that humans used to domesticate plants and animals and results in live forms that have a larger complexity and a winning edge.

We know now that the information that is necessary for the formation of living beings, including humans, is stored in a giant molecule denoted by the acronym DNA, and we know that the DNA of parent life-forms determines the DNA of offspring life-forms. We also know that there are great similarities and connections between the DNA of humans and animals. Yet, the development of human life from more humble beginnings through DNA changes is frequently debated, and some teach the direct intervention of god to create humans and other life-forms. Of course, if we look at the complexities of human life and the DNA, we may well exclaim, as Boltzmann did, "Was it a god who wrote these signs?" However, in 2010, DNA has been artificially created in a biochemistry laboratory of the Venter Institute. The scientists implanted this completely artificial DNA (of a bacterium) into a host cell without DNA. The host cell was awakened to life by this implantation and began to grow and reproduce. We have therefore arrived, at least with some forms of DNA, at the point at which we can experiment at will and predict the outcomes of these experiments with great certainty. We have a scientific understanding of DNA!

Will that end discussions about evolution? Probably not! It is unlikely that we will be able to experimentally reproduce, in the laboratory and within a short time period, the DNA changes that may have occurred over millions of years and that have led to the current DNA forms. Therefore, some experiments that may be crucial to decide such a debate of creation versus evolution of life may not be possible. Nevertheless, science education must be confined to the methods of science, its deductions, and its predictions. It is the predictions of scientific method that make all the difference, and only theories that predict facts that can be experimentally checked should be taught in science classes. Teachings of creation by an all powerful being, as valid as they may be, explain everything but predict nothing. This was already pointed out by the mathematician Pierre–Simon Laplace to Napoleon in the early eighteen hundreds.

This author believes firmly in the theory of evolution. I am also convinced, however, that STEM teachers must be careful not to overstate their case. They should not claim that the evolution of DNA, millions of years ago, is as well understood and established as the science of falling stones.

In spite of the always present limitations of our knowledge, however, it is clear that science provides an extremely useful tool to pursue many great goals of mankind. Science provides also the foundation for engineering methods and enables us to develop technologies that provide us with energy, housing, medication, and recreation, all based on scientific *prediction*. Science thus helps us in our pursuit of happiness. This is what STEM is all about; at least this is what it should be about.

2.1 Physics: Force, Velocity, and Energy

The name physics is closely associated with concepts such as force, velocity, energy, particles, and waves. We will explain these concepts, and what they do for us, in an approximately historic fashion and start with the concept of force.

2.1.1 Force and Energy

The concept of force is probably as old as mankind. Force is necessary to protect yourself from nature, from dangerous animals, from storms by building some kind of roof, to fight enemies, and even to just lift a stone. One certainly thinks in all these connections of muscular force, such as the strength of our arms to lift objects or of our legs to run. The early cultures surprise us often by great buildings such as the pyramids and structures such as Stone-henge, because it is difficult to imagine that such structures were built with the forces that we have normally at our disposal. They show that it must be possible to exert much larger forces than usually attributed to a biceps. To explore these possibilities, we need to understand the nature of what we call force in greater detail.

Some of the features of the concept of force are revealed in Fig. 2.1. This figure shows a seesaw with different board length and weights on each side. The long board with length L_2 needs fewer persons to balance the persons on the short board with length L_1. The persons sitting on the boards exert a force because gravity pulls them down. The important rule for the balance of the forces on each side is

$$F_1 L_1 = F_2 L_2. \tag{2.1}$$

We note that the product of force and length has to be equal on each side to balance the seesaw. We will see below that this fact can be derived from one of the most basic laws of physics, the law of conservation of energy. For now, however, we are

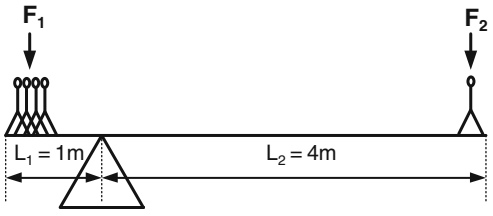


Fig. 2.1 A seesaw having different board length on its two sides. One side has a board length L_2 that is four times as long as the other that has length L_1: $L_2 = 4\,L_1$. One finds for this case that one person on the L_2 side can balance four equally heavy persons on the L_1 side. Because of gravity, the persons sitting on the board exert forces F_1 and F_2, respectively, as indicated, and the seesaw is balanced if $F_2 = 4\,F_1$

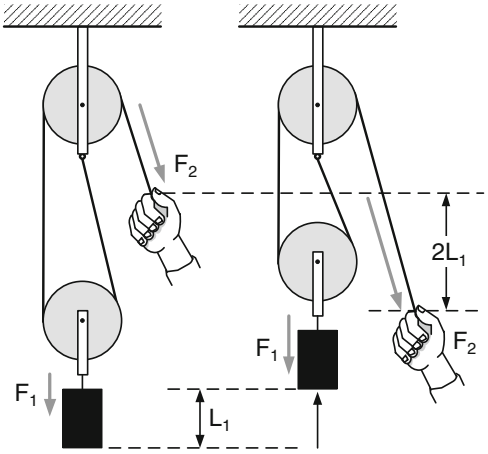


Fig. 2.2 Pulley hoist consisting of two wheels that can rotate around their axis. The axis of one wheel is fixed, for example, mounted to the ceiling, and the second wheel is connected to the first by a rope that winds around both wheels (if n discs are involved around n discs). The weight is pulled down by gravity and thus exerts the force F_1. It can be pulled up by the force $F_2 = \frac{F_1}{2}$ or for n wheels by a force $F_2 = \frac{F_1}{n}$. The ratio $\frac{1}{2}$ of the distance L_1 to the pulling distance $2L_1$ becomes intuitively clear if F_2 points vertically downwards. It stays the same, however, if the force F_2 points in any direction

only interested in the forces that we can exert and how we can increase these forces. We can change the length of the boards so that $\frac{L_2}{L_1} = 10$. Then we need to exert 10 times the force on the short side compared to the force on the long side. This principle tells us why one can move huge stones with a crowbar. The same principle explains why we can exert large forces with wrenches. In fact, many tools of any household are just based on this principle.

Another example is presented by a pulley hoist as shown in Fig. 2.2. This is a well-known tool that can be used to pull up heavy loads. To pull up a load that exerts a force F_1 over a distance L_1, one needs to pull the rope with a force F_2 over a distance L_2. One finds from the experiments that

$$F_2 = \frac{F_1}{2} \tag{2.2}$$

and

$$L_2 = 2\,L_1. \tag{2.3}$$

Taking the product of the left- and right-hand sides of the two equations gives again Eq. (2.1). If we pull up a load with just a fraction of the force necessary to pull it directly, then we have to pull the rope a longer distance corresponding to this fraction. This is universally so for all experimental arrangements; we can think of hoists with more than two wheels that can pull heavier loads. It took a long time until it was understood, that the product of force (in the direction of the movement) and

the distance (that one actually moves) corresponds to the energy that one invests in the process. Energy was up to then, for a variety of reasons, not clearly defined. For example, if it is very hot when persons are pulling loads, it may appear to them that they need a lot more energy in order to accomplish the same task. Physics does not deal with feelings, and based on the rule of Eq. (2.1), as well as other experimental facts, energy is defined by

$$E(\text{Energy}) = F(\text{Force}) \cdot L(\text{Distance}). \tag{2.4}$$

The distance is the length over which the force is applied. If the force changes during the pulling or moving, then one must take the sum of all the distances multiplied by the different forces in the direction of the movement. In the limit of changes over very small distances, this sum needs to be performed according to the rules of Sect. 5.2.2.

Energy is a very important quantity and represents the single most important concept of physics. No machinery, no matter how ingenious, can create energy out of nothing or destroy energy into nothing. This is a simple formulation of the most basic law of physics that we call the law of energy conservation. This law extends beyond mechanical machines and is of general validity. The interested reader should consult the Internet. For mechanical machines energy conservation means the following: we cannot construct any machine, with seesaws, hoists, or whatever, that creates energy. All we can increase (or decrease) is the force. The search for a mechanical machine, that gains energy out of nothing, has been the object of the lifework of many people who wished to create a "perpetuum mobile," a machine that perpetually turns wheels and produces energy. Such a machine cannot be built, because it would violate physic's most basic law, the law of energy conservation. You might ask: how do we know that for sure? Euclid's fifth axiom was not correct, and maybe energy conservation is not correct either! To such an objection one can only say that the law of energy conservation has been tested like no other law of nature, and it has always been found correct.

Forces are not just numbers that characterize a magnitude. The direction of a force is also of great importance. Quantities that are characterized by both magnitude and direction are mathematically represented by "vectors." Like numbers, the mathematical abstraction of a vector follows certain axioms. We will not discuss these axioms here, but rather highlight the main properties of vectors and their application by the following example. Figure 2.3 shows a person pushing a cart on a street or plane that is inclined by an angle α. The force that the person needs to push the cart up the street points in the uphill direction parallel to the street. This force depends, therefore, on the angle α, i.e., on how steeply uphill the cart needs to be pushed. If $\alpha = 0$ then we need practically no force at all to push the cart. As α increases one needs a larger and larger force to push the cart forward. If we wish to push a large weight up a hill and use only little force, then we need to have a small angle α. The architects of the pyramids knew this. They built streets toward the pyramids that had indeed a small angle α of inclination and thus permitted them

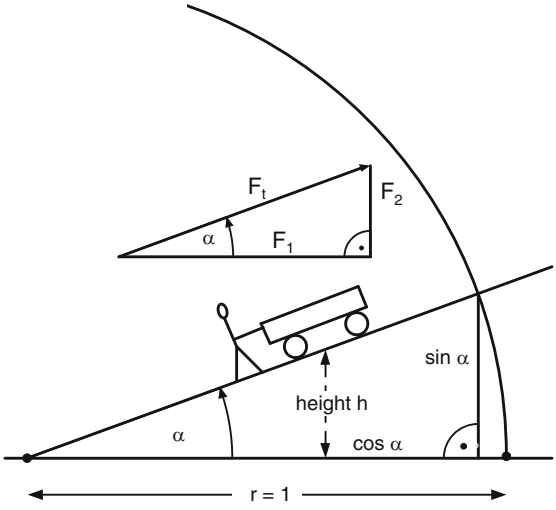


Fig. 2.3 Pushing a cart up an inclined plane illustrates how the total force F_t (that acts on the cart in the direction of the inclined plane) can be seen as being composed of a horizontal force F_1 and a vertical force F_2

to push the big stones upwards to the top of the pyramids. This also illustrates that the force is a vector because its direction is important, and, in our case, the direction also determines how much force we need, i.e., the magnitude of the force.

All of these facts can be understood from the law of energy conservation. Because a horizontal movement does not need any energy (if we forget about the friction of the cart on the ground), all the energy is needed for pushing the cart to a greater height against the forces of gravity. This upward pushing is accomplished by the magnitude of the upwards pointing force F_2 shown in Fig. 2.3. The energy E that one needs to push the cart to a height h is given by

$$E = F_2 \cdot h, \tag{2.5}$$

which explains why we can do the work with a small force F_2, if we have a small inclination α of the street. The exact geometrical and mathematical relationships of the forces that are involved in the example of Fig. 2.3 can be derived as follows. We have drawn in the figure a circle with a radius of unit length (e.g., one meter) and have also indicated a triangle. This triangle includes one right angle and the angle α, and its longest side is equal to the radius of the circle. The lengths of the other sides are equal to $\sin(\alpha)$ (vertical) and $\cos(\alpha)$ (horizontal), respectively. This triangle is similar (in the mathematical sense) to the triangle that shows the forces. From the rules given in Sect. 1.3.1 for similar triangles, we obtain then the following relation for the magnitudes of the forces:

$$F_2 = F_t \sin(\alpha) \tag{2.6}$$

and

$$F_1 = F_t \cos(\alpha). \tag{2.7}$$

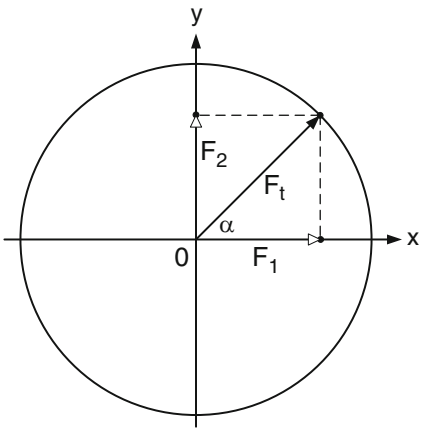


Fig. 2.4 Radius vectors are defined as the *directed line* going from the origin of the coordinate system to any point. The fact that these vectors indicate a direction is expressed by the *tip of an arrow* at the endpoint of the line

This decomposition of the total force F_t into a horizontal component F_1 and a vertical component F_2 is useful because it helps us to calculate the force that is necessary to do the work. The work equals the energy that is needed. This way of dealing with forces is most efficiently done with the mathematical concept of a vector. It is convenient to give the explanation of vectors by using a Cartesian coordinate system and defining a "radius vector" following the illustrations of Figs. 2.3 and 2.4.

Radius vectors are the directed lines that start from the origin of the coordinate system and extend to any point (x, y) of the plane of the coordinate system. Radius vectors are usually denoted by bold-faced letters like $\mathbf{r}$ and are characterized and represented by the coordinates (x, y) of the point; thus we have $\mathbf{r} = (x, y)$. The vector of Fig. 2.3 that points from the center of the circle in the direction of the inclined plane toward the point that intersects the circle has the coordinates $(\cos(\alpha), \sin(\alpha))$ and, therefore, $\mathbf{r} = (\cos(\alpha), \sin(\alpha))$. Such a vector provides us with the length and the direction of a line. The physical sciences deal with many different vectors. As we just have learned, the force is a vector. Therefore electric and magnetic fields are also vectors because they are forces that have a direction and a magnitude. How do we deal with such general radius vectors? Exactly the same way as we deal with vectors that represent a length and a direction! There is just a little trick necessary. We plot the coordinate system with a line pointing from the 0 to the endpoint of the vector. However, now, we do not deal anymore with distances but with forces. We therefore just replace the unit distances on the x, y axis by the unit forces. Figure 2.3 shows the result. The vector from the 0 point of the coordinate system to the circle has now the magnitude F_t, and we call it the vector $\mathbf{F}_t$. Thus we only need to multiply everything by the length of the vector which in the above example is the magnitude of the force. The vector $\mathbf{F}_1$ has the coordinates $F_t(\cos(\alpha), 0)$ and the vector $\mathbf{F}_2$ has coordinates $(0, F_t \sin(\alpha))$. We can therefore write

$$\mathbf{F}_t = (F_t \cos(\alpha), F_t \sin(\alpha)),\ \mathbf{F}_1 = (F_t \cos(\alpha), 0),\ \mathbf{F}_2 = (0, F_t \sin(\alpha)). \qquad (2.8)$$

We define then the addition of vectors in the following way. Geometrically speaking, we add vectors $\mathbf{F}_1$ and $\mathbf{F}_2$ by putting the lower end of vector $\mathbf{F}_2$ at the right end of the vector $\mathbf{F}_1$, and the vector $\mathbf{F}_t$ is obtained exactly that way as shown in Fig. 2.4. Algebraically this means vectors are added by just adding the x- and y-components separately. Thus the vector addition

$$\mathbf{F}_t = \mathbf{F}_1 + \mathbf{F}_2 \tag{2.9}$$

is for the radius vectors equivalent to

$$(F_t \cos(\alpha), F_t \sin(\alpha)) = (F_t \cos(\alpha), 0) + (0, F_t \sin(\alpha)), \tag{2.10}$$

which represents the algebraic way to add vectors. Note that F_t is just the magnitude or length of the vector $\mathbf{F}_t$. It is a good exercise to add vectors both in a geometrical fashion as shown above with the cart and algebraically as just explained.

In this way we can add forces that point in arbitrary directions and obtain the total resulting force. This rather involved procedure of adding vectors is of great importance in physics. As mentioned, many physical quantities are vectors, because they are involving not only a magnitude but also a direction. The velocity of physical objects is also a vector and so is the acceleration that we will define later. For someone who likes physics, it is therefore very important to get familiar with the concept of vectors. However, in most of what follows, we have avoided this rather sophisticated type of calculation, by assuming to start with that only one direction, for example, the x-direction, is relevant.

2.1.2 Momentum: The Mechanics of Billiards

We discussed in Sect. 2.1 the very important law of energy conservation. There is a second law for the mechanics of objects that is related to energy conservation and also of great importance. This is the law of conservation of momentum. Momentum is, in contrast to energy, a vector, i.e., a quantity for which direction is of importance and is denoted by a bold letter, for example, by $\mathbf{p}$. Thus we have a symbol for the word momentum, but what is it and what is conserved? Momentum is related to both the velocity of an object (also a vector) and to the energy of the object as illustrated in Fig. 2.5. We therefore discuss first the definition of velocity.

The concept of velocity is well known from our daily life. The velocity of an object is obtained by dividing the total distance that the object travels by the time that it takes for the travel. Consider an object that travels in the x-direction of a coordinate system. Then, if the object starts at the point x_1 at time t_1 and arrives at point x_2 at time t_2, we calculate the velocity v from

$$v = \frac{x_2 - x_1}{t_2 - t_1}. \tag{2.11}$$

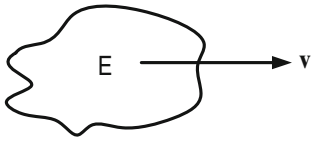


Fig. 2.5 A physical object with total energy content E moves with velocity **v**. The velocity is a vector and therefore denoted by the bold-faced **v**. The "momentum" is also a vector and is by definition proportional to the velocity. The constant of proportionality is described in the text

It is customary to denote the differences $x_2 - x_1$ by Δx and $t_2 - t_1$ by Δt, respectively. Thus we have

$$v = \frac{\Delta x}{\Delta t}. \tag{2.12}$$

For the case of very small differences Δ (see Sect. 5.2.1 for a detailed explanation), we write

$$v = \frac{\mathrm{d}x}{\mathrm{d}t}. \tag{2.13}$$

Generally the velocity is a vector **v** in two or three dimensions, i.e., in a plane or in space, respectively.

Even if we consider only the x-direction, the velocity is not a simple number. The velocity of a car is, for example, given in kilometers (or miles) per hour. Other units can also be used. If we wish to have Δx in meters and Δt in seconds, then we obtain the velocity in meters per second. Units like this are important for science and engineering problems. The use of such units usually presents some problems to students, because there is a new concept involved here. When we talked about adding and subtracting numbers, we stated that we can substitute all kinds of things for the numbers, for example, apples and oranges. We only needed to make sure that we add and subtract only the same kind, i.e., either apples or oranges. Clearly it makes no sense to subtract five apples from ten oranges. Physical quantities such as the velocity are composed of both space and time measurements, and the velocity is obtained by dividing distances by times. Therefore, the unit for the velocity is composed of two different separate units, for example, of the units kilometer and hour. We can still add and subtract such quantities. However, we need to make sure that all these quantities are given in the same units. It does not make any sense to add or subtract kilometers per hour from meters per second.

We return now to momentum and define the momentum as a vector **p** that is proportional to the vector of the velocity **v** and also proportional to the total energy content E of the object. The object is shown in Fig. 2.5 as a "blob" with arbitrary shape symbolizing that it may consist of a single particle, such as an electron or of a complicated collection of many bodies that interact with each other by the forces of nature such as a molecule. It may also be a big object like a billiard ball or even a planet.

Thus we have the definition

$$\mathbf{p} := KE\mathbf{v}. \tag{2.14}$$

Proportional means equal except for a constant that we have named K in this case. The symbol := means that this equality is valid by definition and does not represent by itself any law or rule. This particular definition of momentum encompasses the work and results of many famous physicist over many centuries. It is therefore not obvious why we have defined "momentum" that way and what benefits this definition gives us. It was Einstein who realized the full meaning of the concept of momentum. Galileo and Newton had already realized the following. If that "blob" of energy that we have shown in Fig. 2.5 moves somewhere out in space and no force acts on it, then it will move on forever with the same velocity. Thus, the momentum $\mathbf{p}$ is a quantity that is conserved and does not change if we do not apply any forces. Einstein also found the constant K from his theory of relativity that is explained in Sect. 5.1. He found that

$$K = \frac{1}{c^2}, \tag{2.15}$$

where $c = 300{,}000\,\text{km/s}$ is the velocity of light. As we have explained also in Sect. 5.1, Einstein derived the very famous relation

$$M = \frac{E}{c^2}, \tag{2.16}$$

where M is the mass of the blob that we could measure if we would try to accelerate or weigh it. Therefore we have

$$\mathbf{p} := M\mathbf{v}. \tag{2.17}$$

This equation was the definition of Newton. Galileo and Newton were the first to realize the importance of the following law of physics: if an object moves and there are no forces acting on it, then the object will move on with the same momentum forever. There is a great deal of abstraction here, because objects usually are influenced by forces. Gravity acts on all objects on earth and most objects are subject to a force of friction that tends to stop the motion by creating heat energy (see Sect. 2.3). Even far out in space there is usually some gravitational force, for example, that of the sun, acting on objects. So Newton's abstraction was a big step and not obvious at all. But why is this law so important? This is discussed in the following by using the examples of the movement of billiard balls, rocket engines, jet engines, and other mechanical phenomena.

Billiard balls are made out of a fine material that has virtually no friction, neither with the billiard table nor with other billiard balls. Gravity acts on them. However, because the billiard balls are on a very horizontal table, the gravitational forces are canceled out by the table that pushes against the balls. Therefore on a billiard table we can observe Newton's famous law by moving a billiard ball: it moves on with the same speed as long as no forces act on it, meaning as long as it does not hit something. What if a billiard ball hits centrally another one that stands still?

You might think that somehow both will then move on. This is not so! The ball that hits stops entirely and the other one moves on with the same speed that the first ball had. The reason is that both momentum and energy of both billiard balls need to be conserved, and this is only possible that way.

If we push the billiard ball over a distance, that means we exert a force on that ball over that distance. As we know, force times distance equals the energy that we supply. All that the billiard ball does, as a consequence of this energy supply, is that it moves with a certain velocity. One therefore says that one has changed the energy supplied by the push into moving energy or "kinetic energy." We describe now a few properties of this kinetic energy and then give the equation from which one can calculate it easily. How can one measure kinetic energy? There are, of course, many elegant ways to do so. We mention here a very inelegant but very important way. Consider a car that moves with a certain velocity v. If the car smashes into a wall, then the kinetic energy goes to zero because the car is stopped. Energy cannot be destroyed, so where is it? If you look at the remnants of the car then you see that steel has been distorted, windows have been smashed, and other damage has occurred. Clearly you need energy (forces over a distance) for all of this. Also, and this is very important, the parts of the car that have been severely distorted have heated up. Thus a lot of energy has been transformed into heat (see Sect. 2.3). If the car would not have driven into a wall, but stopped by use of the brakes, then all the energy would have been transformed into heat at the brakes. The brakes are therefore heating up when used. If you use the brakes of a car too frequently, for example, when going downhill, then the brakes will get very hot and may start burning. The same problem happens, if you forget that the brakes are on while driving. Thus one can measure the kinetic energy by just measuring the heat that is produced when braking or the damage plus heat that is generated when smashing the car. Modern cars, so-called hybrids, can turn the braking energy into electrical energy that can be used again. The kinetic energy is thus turned into electrical energy that also can be measured. This would be a more elegant way of measuring kinetic energy.

We turn now to the details of the rules for the kinetic energy. We know that there is more damage to the smashed car that drives into a wall with higher velocity v. Careful measurements of the damage and heat generation for a given velocity show the following. If we double the velocity, the damage and heat is not just twice as large but four times as large. If we triple the velocity the damage and heat is nine times as large. You can guess that for four times the velocity we have sixteen times the damage and heat. Thus the kinetic energy that we denote by E_{kin} increases with the square of the velocity. One also finds that the kinetic energy increases with the mass of the car. A truck can cause much more damage than a small car. Here one finds that if one doubles the mass the damage doubles, if one triples it, the damage triples, and so forth. Thus the kinetic energy is proportional to the mass. As we will see later, for the ordinary velocities that we deal with on earth, the factor of proportionality is just $\frac{1}{2}$ and we have

$$E_{\text{kin}} = \frac{1}{2} M v^2. \tag{2.18}$$

Consider now the motion of two billiard balls when they collide. We start with one ball moving with velocity v_1 that hits a second ball centrally. The second ball is standing still and has therefore velocity $v_2 = 0$. After the collision we denote the velocities by v_1' and v_2', respectively. Both billiard balls have the same mass M. Because momentum is conserved, meaning that the total momentum before and after the collision is the same, we have

$$M v_1 = M v_1' + M v_2'. \tag{2.19}$$

Canceling M on both sides gives

$$v_1 = v_1' + v_2', \tag{2.20}$$

and squaring left and right side of this equation we obtain

$$v_1^2 = v_1'^2 + 2v_1' v_2' + v_2'^2. \tag{2.21}$$

From Eq. (2.18), the law of conservation of energy, we get, after again canceling out the equal mass

$$v_1^2 = v_1'^2 + v_2'^2. \tag{2.22}$$

Subtracting now Eq. (2.22) from Eq. (2.21) we have

$$0 = 2v_1' v_2' \tag{2.23}$$

which means that either v_1' or v_2' or both need to be 0. Because both cannot be zero (that would mean the energy has vanished), the only solution that makes sense is that $v_1' = 0$. Then we obtain from Eq. (2.20) the result $v_2' = v_1$. This means that the first ball stands still and the second moves away with exactly the same velocity that the first had. This is a well-known result and is often shown in experiments. You can try it also yourself on a billiard table. The experiment becomes even more astonishing if one does the experiment with three balls as shown in Fig. 2.6. There is a common toy found in select gift shops. Several stainless steel spheres are hanging on a string next to each other. If one takes the first one and lets it swing so that it hits the others, only the last one flies off. When this last one returns and hits, the first one swings back and so forth.

2.1.3 *Acceleration: The Mechanics of Falling Stones and Planets*

The title of this section may strike the reader as strange. "Falling stones and planets" sounds like planets are falling exactly as stones do. This is no mistake! The laws of falling stones and planets orbiting the sun are really very similar, almost identical, and it was Newton who had the great idea that the orbiting of the planets can be

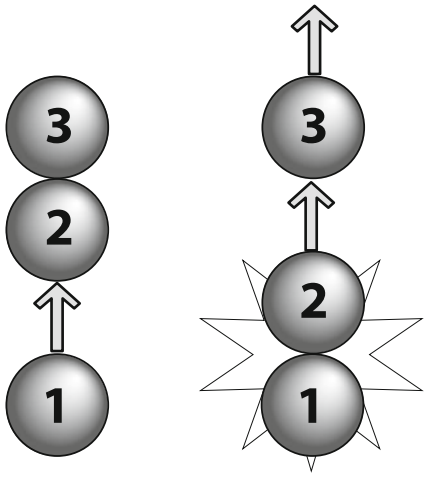


Fig. 2.6 *Left*: billiard ball 1 hits balls (2, 3) *centrally*. The result is shown to the *right*: balls (1, 2) stop and (3) moves on. This is a demonstration of the conservation of both momentum and energy

understood the same way as the falling of stones. Imagine that you are standing at the shore of the ocean and that you are throwing a stone toward the horizon. The stone starts to move horizontally but then is attracted by the gravity of the earth and finally falls into the sea. Now assume that you are throwing the stone faster. Then it moves further out before dropping into the water. Now push your fantasy a bit further and assume that you can throw the stone so fast that it drops just as much as the earth (and therefore the water) curves, because the earth is round. What you have then is a stone that never falls in the sea, it moves forever like a satellite. Of course, the air friction would slow the stone down, and it would eventually fall into the ocean. However, if we perform the experiment very high up in space, then there is no air, and we really would have a satellite! Naturally, Newton did not think of satellites. He just thought of the moon in exactly the same way, as if it were a stone thrown with high speed and circling the earth. Newton calculated the moons orbit that way, and it came out about right. We discuss this great idea below, starting with the laws of falling stones that were already discovered by Galileo well before Newton.

According to legend Galileo dropped stones from the leaning tower of Pisa and observed them carefully. He recorded how fast they were falling and hitting the ground. Galileo actually performed quantitative experiments with balls rolling down a ramp. One of Galileo's important notions was an idealization. He thought that if one could remove the air and create a vacuum, all things fall equally fast. This is indeed true and has been proven by now beyond any doubt. In air, down feathers fall to the ground much slower than a piece of metal. This is due to the interaction of the feather with the air. A little wind might even lift the feather up to higher elevations. A heavy steel ball on the other hand is not much impeded by air or influenced by wind. Different kinds of steel balls fall also all equally fast. Galileo's theory that all things fall equally fast in vacuum has been fully confirmed

by experiments. Galileo found from his experiments that the time period t that an object falls (starting at $t = 0$) is related to the height h that it falls by

$$h = \frac{gt^2}{2}. \tag{2.24}$$

Galileo also found that the velocity v of an object, that is standing still at $t = 0$ and starts then falling while we measure the time t, obeys the equation

$$v = gt. \tag{2.25}$$

Here g is a constant that represents the acceleration of an object by the attraction (gravity) of the earth. As an aside, note that these laws are the laws of nature that underlie the computer simulations of the game "crazy birds." It is a nice class project to calculate the trajectory of a thrown stone.

In general, acceleration is denoted by the letter "a" and is defined as the change of velocity with time:

$$a = \frac{\Delta v}{\Delta t}. \tag{2.26}$$

Here $\Delta v = v_{\text{after}} - v_{\text{before}}$ and $\Delta t = t_2 - t_1$, where v_{after} is the velocity at the time t_2 after the acceleration and v_{before} is that at the time t_1 before the acceleration occurred. The change of physical quantities is usually denoted by the Greek symbol Δ, pronounced "delta." In the limit of very small changes, Δ is replaced just by the letter "d" that denotes the so-called "differential" and is explained in more detail in Sect. 5.2.1. By this convention, the acceleration is written as

$$a = \frac{\mathrm{d}v}{\mathrm{d}t}. \tag{2.27}$$

For example, if an object changes its velocity during the time period of 2 s from a velocity of one meter per second to three meters per second, then $\Delta v = 3 - 1 = 2\,\mathrm{m/s}$ or $2\,\frac{\mathrm{m}}{\mathrm{s}}$ and $\Delta t = 2\,\mathrm{s}$. The acceleration a of this object is then $a = \frac{\Delta v}{\Delta t} = 1$ meter per second squared or $a = 1\,\frac{\mathrm{m}}{\mathrm{s}^2}$. Note that the velocity is given in meters per second, and the acceleration is obtained from the velocity change by another division by the time difference. Therefore the acceleration must be measured in meters per second squared or $\frac{\mathrm{m}}{\mathrm{s}^2}$ (or kilometers per hour squared, etc.).

For the acceleration g of any object by the attraction of the earth, Galileo found experimentally that

$$g = 9.81\,\frac{\mathrm{m}}{\mathrm{s}^2}. \tag{2.28}$$

What does this mean? Lets assume a stone falls for 10 s. Then, according to Eq. (2.24), the height it falls is $9.81\,\frac{\mathrm{m}}{\mathrm{s}^2}100\,\mathrm{s}^2/2 \approx 490\,\mathrm{m}$. The stone hits the ground with a velocity of $v = 9.81\,\frac{\mathrm{m}}{\mathrm{s}^2}10\,\mathrm{s} = 98.1\,\frac{\mathrm{m}}{\mathrm{s}}$ according to Eq. (2.25). The velocity given in meters per second can easily be converted into kilometers per hour (try this as a little homework problem), and one finds the value of about 353 km/h. This is

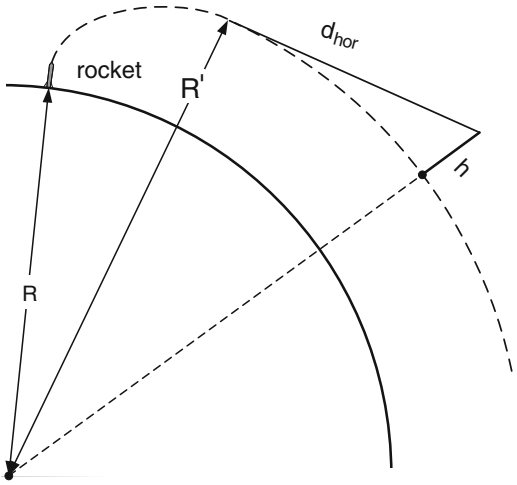


Fig. 2.7 The illustration shows the surface of the earth (radius R) and a rocket carrying a satellite into a circular orbit with radius R'. Note the tangent to the circle of the orbit that is drawn over a distance d_{hor} as well as the height h that is also indicated. These two lines are two sides of a triangle that we also discussed in connection with Fig. 1.19

the speed of an airplane, and here we have the reason for the fact that falling from a height of 490 m is absolutely destructive. The equations found by Galileo are interesting and can be applied to many problems of our daily life. However, their impact was bigger than that.

Newton realized that one could use these laws also to calculate the orbit of planets and moons. To understand this we present Newton's idea, not by discussing planets or moons but by calculating the orbit of a satellite. Newton, of course, did not know what a satellite was or could be. The first satellite (Sputnik) was launched in October 1957, 230 years after Newton died. However, Newton's ideas were sufficient to calculate how to launch such a satellite. Figure 2.7 shows a rocket starting from earth (that has a radius R) and arriving at a certain orbit (that has a radius R'). At that point of entering the orbit, the rocket and its satellite begin to move perpendicular to the radius of the orbit along the tangent of the circle. We have included in the figure a triangle (on top of the satellite orbit) that is basically identical to the corresponding triangle in Fig. 1.19 of Sect. 1.3.3. This triangle has one side with a height h and another which is the distance to the horizon d_{hor}. The triangle in Fig. 2.7 follows the same geometry laws and relations except that the radius of the earth R must now be replaced by the radius of the satellite orbit R'. We therefore obtain from Eq. (1.82)

$$d_{\text{hor}} = \sqrt{2hR'}. \tag{2.29}$$

Newton's great idea was now the following. Assume that the satellite moves so fast that, while moving the distance that is denoted by d_{hor}, it "falls" precisely by the amount h; that means, it falls exactly the amount of the curvature of the orbit. Then the satellite "falls down" but actually keeps always the same distance R' from the surface of the earth. Is this possible? We use now Galileo's laws of falling stones and will show that it is possible indeed. First we use the definition of velocity,

which equals to the distance traveled by the object divided by the time needed for traveling that distance. Therefore, if the velocity of the satellite is v, then the time needed to travel the distance d_{hor} is

$$t = \frac{d_{\text{hor}}}{v}, \tag{2.30}$$

and using Eq. (2.29) one obtains

$$t = \frac{d_{\text{hor}}}{v} = \frac{\sqrt{2R'h}}{v} \tag{2.31}$$

of which we take the square to have

$$t^2 = \frac{2R'h}{v^2}. \tag{2.32}$$

We have thus obtained the square of the time that it takes the satellite to travel the distance d_{hor}. In that time, we wish that the satellite falls downwards by exactly the height h. We know from Galileo's law of Eq. (2.24) that the square of that time must be

$$t^2 = \frac{2h}{g}. \tag{2.33}$$

As one can see by comparing the last two equations (2.32) and (2.33), the satellite falls exactly the height h if

$$v^2 = gR'. \tag{2.34}$$

This means that a satellite with the precise velocity of

$$v = \sqrt{gR'} \tag{2.35}$$

will circle the earth basically forever and not fall down to the surface. Therefore, if we wish a satellite to orbit the earth at a distance of $R' = 7{,}000\,\text{km}$ or equivalently $R' = 7 \cdot 10^6\,\text{m}$ (where m stands for meters), then we need a velocity that satisfies

$$v^2 = 9.81\,\frac{\text{m}}{\text{s}^2} 7 \cdot 10^6\,\text{m} = 7.87 \cdot 10^7\,\frac{\text{m}^2}{\text{s}^2}. \tag{2.36}$$

Taking the square root of this equation on both sides we obtain a velocity $v = 8.29 \cdot 10^3\,\frac{\text{m}}{\text{s}}$ or $8.29\,\text{km/s}$. This is a very high velocity, but it is achievable with rockets, and this is how Russia launched the first satellite with the name Sputnik.

Newton did, of course, no such calculation for a satellite. He did it for the moon! To perform the same orbital calculation for the velocity of the moon, he needed to figure out two more laws that certified him as one of the greatest physicists of all times. Newton studied the astronomical data and corresponding rules that were derived by Kepler. He realized then, by looking at Kepler's data and Kepler's so-called third law (the proportionality of the square of the orbiting time to the third

power of the orbiting radius), that the gravitational force of the earth (or sun) must be weaker the farther we go away from the earth (or sun), and that it decreases with the square of the distance. This is the major finding that we must remember to understand Newton's calculation of the moons orbit. Thus, he knew that he could not use the value $g = 9.81\,\frac{\mathrm{m}}{\mathrm{s}^2}$ for the falling acceleration of the moon, but he had to reduce g with the square of the distance from the earth.

Newton actually deduced from the astronomical data the full law that tells us the magnitude of the gravitational attraction between two objects, and we state this law here for completeness. The gravitational force F by which two objects such as the earth with mass M_{ea} and the moon with mass M_{mo} attract each other is

$$F = k_N \frac{M_{\mathrm{ea}} M_{\mathrm{mo}}}{\mathrm{d}^2}, \tag{2.37}$$

where k_N is called the gravitational constant. The gravitational constant can be expressed in units of $\mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$. In words, this reads meters to the third per kilogram and per square seconds. The value in these units is $k_N = 6.673 \cdot 10^{-11}\,\mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$. Note the decrease of this force with the square of the distance d. We need now a connection between force and acceleration to obtain the precise relation between acceleration and distance of the objects in question. Newton found this law that is also needed to calculate the Moon's orbit. It is Newton's most famous physics law and gives the connection between any force F and the acceleration a of an object with mass M that is subjected to this force:

$$F = Ma. \tag{2.38}$$

Knowing that the acceleration of the moon by the gravity of the earth decreases with the square of the distance of the moon from the earth, Newton calculated the orbit of the moon as follows. The distance to the moon is 60 times the radius of the earth, and Newton knew this from Kepler's third law. Therefore we can calculate the acceleration a with which the moon "falls" by dividing g by $60^2 = 3{,}600$. We know this from Eqs. (2.37) and (2.38). Then we obtain from Eq. (2.34) the velocity of the moon by replacing R' by $60R$ where $R = 6.37 \cdot 10^6\,\mathrm{m}$. This gives for the velocity of the moon $v = 1{,}020\,\frac{\mathrm{m}}{\mathrm{s}}$. The length of the moon orbit is $2\pi R' = 2\pi 60R = 2.40 \cdot 10^9\,\mathrm{m}$. The time the moon needs then to go full circle is obtained by dividing this orbital length by the moons velocity, which gives $2.35 \cdot 10^6\,\mathrm{s}$. This is about 27.25 days, which is very close to the moons orbiting time.

The precise calculation for the moon is more difficult and was a problem even for Leonhard Euler, one of the greatest mathematicians of all time. The reason is simply that not only the earth attracts the moon but also the sun. Therefore one deals with three objects, sun, moon, and earth. This is called a three-body problem. If one wishes to be more precise one even has to include the other planets and ends up with a many-body problem. These problems have no exact solution. However, one can solve them by solving Newton's equations by using a computer. This can be done basically to any desired accuracy and has been done with greatest precision to send astronauts to the moon. Of course, the moon mission that resulted in the first

moon landing on July 20, 1969, required more than this calculation. It did require the technology of the rocket engines that propelled the giant Apollo rocket and the moon lander, and it required the technology to sustain the astronauts in space and to return them safely to earth. When Neil Armstrong made the first step on the moon and looked up to see the earth he said these famous words: "That's one small step for a man, one giant leap for mankind." Indeed it was. Here we can see the symbiosis of science, engineering, and technology in the best light, and we can see the enormous developments that are necessary to bring the idea for the laws of motion in space to the first step of an astronaut on the moon.

Of course, the orbits of the planets around the sun can be understood from Newton's laws as well. They are more complicated to find than the circular satellite orbits that we calculated above, because the shape of these orbits is ellipses and only almost circles. However, all the calculations for these planetary ellipses follow the same principles that we have just discussed. Details of all of these orbit calculations for planets can be found on the Internet and provide nice student projects, particularly for those who have mastered Sects. 5.2.1 and 5.2.2.

In summary, we have learned in this section, that we can calculate the orbit of a satellite just by applying Galileo's laws for falling stones. The calculation of the orbit of the moon requires also the knowledge of how gravity diminishes with distance. Newton provided these laws that give us the understanding of massive objects orbiting around each other.

2.1.4 Spinning Objects: From Gyros to Electrons

Spinning discs with an axle, also called gyros, never have ceased to amaze me. They fascinated me as a child, as a student, and they fascinate me today, while I write this book. The really astonishing effects can be seen only when the disc is rather heavy and spins very very fast. Then, the fast rotating disc appears to be a totally different object as compared to a nonrotating disc. For example, if the rotating disc has an axle with a thin tip and if we put the tip down on a smooth surface (e.g., a table), the gyro stays upright for a long time, while a nonrotating disc immediately tips over. This is illustrated in Fig. 2.8.

Even more astonishing is the fact that a disc that is specially mounted as shown in Fig. 2.9 does not drop downwards, but circles slowly keeping the vertical distance from the ground almost constant. If the disc does not spin, it drops down immediately.

This looks almost magical, and you have to see real experiments to believe it. Such gyros are, of course, available commercially and hopefully ready for experimentation in every science classroom. Newton concluded from such experiments that the spinning disc behaves different because it spins in space, and this space is at rest compared to the disc. Mach did not believe that and suggested that it was the spinning compared to the other masses in the universe that made the difference.

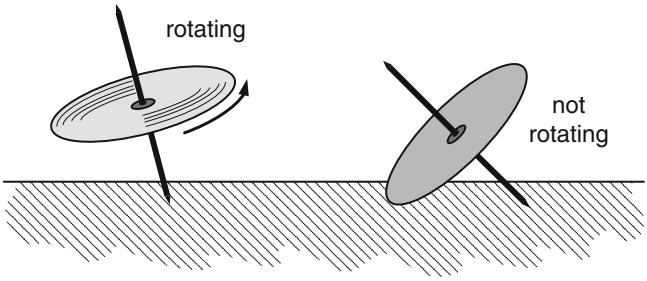
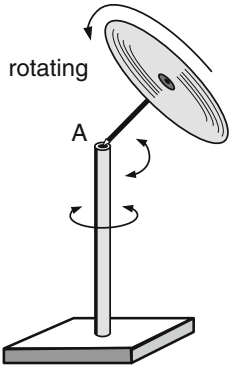


Fig. 2.8 A rotating disc, with an axle that touches a surface with its tip, will not tilt over toward the surface, at least not for a while. A disc that does not rotate will quickly drop down and touch the surface

Fig. 2.9 Spinning disc mounted on an axis that can freely rotate both vertically and horizontally. If the disc spins very fast, it will not fall down but rotate around the central stand



Einstein was at first impressed by Mach's view, but later changed his mind. I think it is safe to say that we really do not know the root cause why spinning discs are behaving so differently to the resting ones. As far as I understand it, this must be linked to what space and vacuum really are, and you can find more about it in Sect. 5.4.

It is a fact, however, that the behavior of spinning objects derives directly from the inertia of moving bodies: a body with mass M that is not subjected to any forces will move on forever with the same velocity because momentum is conserved. This connection of momentum conservation and behavior of spinning discs can be seen from the following two famous experiments.

The first experiment (experiment 1) is performed with a gyroscope. A gyroscope is a spinning disc whose axle is mounted such that it is free to take any orientation. The way this mounting is done is shown in Fig. 2.10. If the disc spins very fast, it does not matter how the outer frame is rotated or moved. The axis of the spinning disc will always point in the same direction. This is why the gyroscope can be used as a compass and tell you the direction you are going, and it is actually often much more reliable than a magnetic compass needle. The direction of the Hubble telescope that is orbiting in space, for example, is controlled by a gyroscope. In

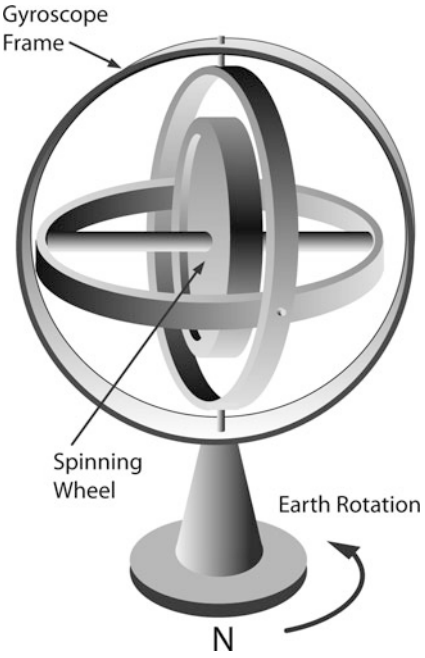


Fig. 2.10 A gyroscope consists in essence of a spinning disc with an axis that is mounted in such a way that it can freely turn in any direction

experiment 1 we put the gyroscope of Fig. 2.10 directly at the north pole, with its stand pointing in the direction of the axis of the earth. One of us has to go to the north pole also and watch the rotating disc of the gyroscope. What that person will see is the axle of the disc performing a horizontal rotation. If the axle points in a certain direction at the start of the experiment, it will do so again after 24 h. If the disc does not spin then the axle does not rotate. What is happening here? Of course the earth has rotated around itself in 24 h and the observing person has rotated with the earth. The spinning disc, however, kept a constant direction and was not influenced by the rotation of the earth. The non-spinning disc, on the other hand, turned just as the observer turned with the earth and therefore appeared to the observer as standing still. What has that to do with inertia and momentum conservation? This is illustrated by the second experiment (experiment 2) that is also performed at the north pole but with a pendulum instead of the gyroscope.

A pendulum is consisting of a heavy object hanging on a rope. For our case it is important to use a long rope (at least a few meters long) and a very heavy object (at least a few kg or pounds). The rope compensates for the action of gravity and holds the heavy object against the gravitational force. We can therefore compare the motion of this object in some respect to the motion of an object free of forces. Such an object will follow the law of momentum conservation and continue to move without changing directions. For the pendulum this is not quite true, because as it swings away from its lowest position, the heavy object goes against gravity and loses kinetic energy until it stands still and then reverses and moves in the

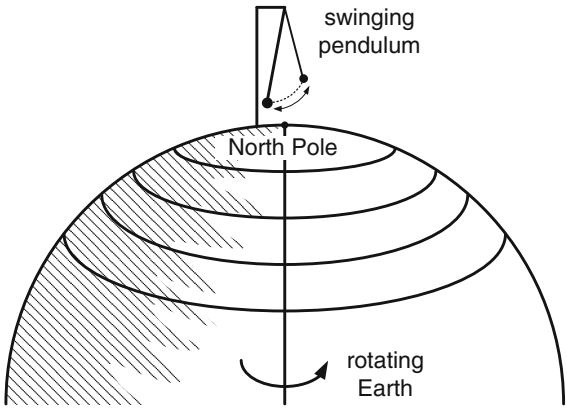


Fig. 2.11 A pendulum swinging at the north pole will swing in a plane. For an observer at the north pole, this plane appears to rotate a full turn in 24 h. It is, of course, the observer and the earth that rotate, while the pendulum swings in the same plane

opposite direction. However, the actual circle along which the heavy object moves stays in the same plane because gravity accelerates only toward the center of the earth. Therefore if we have a pendulum at the north pole, it will swing in one plane and that plane will seem to rotate a full turn in 24 h exactly as the gyroscope. This experiment is illustrated in Fig. 2.11. Of course, it is again the earth that rotates while the pendulum stays in one plane. This similarity between pendulum and gyroscope suggests that the "strange" effects that gyroscopes show are in essence due to the inertia of massive bodies and the law of momentum conservation. To do these two experiments, you actually *need not* go to the north pole. You can see the rotation of the earth clearly in your own home as long as you are not living too close to the equator. The problem is only to have a pendulum with long enough rope and heavy objects so that the pendulum oscillates long enough to see the earth rotation, that is at least for several hours.

The special properties of spinning objects are very important in many subareas of physics. Even electrons, protons, and atoms do have a so-called "spin" and the properties of this spin are reminiscent of those spinning objects. There are, however, important differences between spinning electrons and classical gyroscopes that are discussed in quantum mechanics courses at the university, which provide a complete theory for the electron, proton, and atom spin. This is an advanced topic that was pioneered by P.A.M. Dirac and is discussed in Sect. 5.3.

2.1.5 General Properties of Waves: From Sound to Tsunamis

We know waves from watching water in the wind. The water moves up and down at any given spot showing valleys or minima and mountains or maxima. At the same time the minima and maxima move into some direction with a certain speed also called the velocity of the wave. At any beach you can see that the velocity of the

waves is about a few meters per second. This is the velocity with which a surfer rides with the wave. Ordinary water waves are thus rather slow. Electromagnetic waves have the highest possible velocity, the speed of light, as we will learn in Sect. 5. The speed of waves in water, in solids, or in air depends on how the liquid, solid, or gaseous material that carries the wave actually moves. This movement can occur in different ways even if we have the same material. For example, for ordinary surface water waves, the particles of water move up and down, and the motion is rather slow, as mentioned, a few meters per second.

There are also other possibilities of creating waves in water. A strong earthquake can lead to an enormous compression of water because of a rapid movement of the bottom of the sea. For example, in the earthquake of Japan in March 2011, a large area of the bottom of the sea moved upwards within a very short time. The water cannot react immediately to this rapid movement and is therefore compressed by enormous forces. We know that compression and subsequent dilatation of air gives rise to the ordinary sound waves. These travel with a speed of 343 m/s (referred to as Mach 1). Pressure-dilatation waves also propagate in water. The water-pressure wave after an earthquake moves mostly below the water surface. It is much faster than the ordinary water waves at the surface and can be several hundreds of kilometers per hour. The actual speed, which matters in catastrophes called Tsunamis that often follow earthquakes at the bottom of the sea, can be determined very precisely by computer simulations. When the fast-moving pressure-dilatation waves approach the shorelines, they convert their enormous energy into surface water waves of ordinary speed (around 20 km/h). These ordinary waves may then be of vicious duration and power, depending on the magnitude of the earthquake. The destructive Tsunami of the 2011 Japan quake destroyed Japanese shorelines and cities and still caused powerful destruction, when the pressure wave arrived about 8 h later, thousands of miles away in Hawaii. In addition, and this is an important part of the destructive action of Tsunamis, they have very long wavelength. Therefore the first sign of a Tsunami is often that the water leaves the coastlines as if there would be very low tide. Then the water returns, and because of the long wavelength returns, and returns, and returns.

Mathematical Description of Waves

There are three important values that are characteristic for waves: the velocity v, the frequency ν, and the wavelength λ of a wave. We have discussed already some aspects of the velocity. The frequency is defined as the inverse time period that a wave needs to complete one full cycle. Consider, for example, a water wave that has a maximum height at a certain location and time. Then the water moves downwards and upwards again until it reaches again the maximum. If this process takes 5 s then the frequency is given by $\nu = \frac{1}{5\,\mathrm{s}} = \frac{1}{5}\,\mathrm{s}^{-1}$. The frequency of waves can vary in wide limits and may be extremely high for some electromagnetic waves, as we will see below. The wavelength is the distance from one wave maximum to the next maximum, or from one wave minimum to the next, and is usually denoted by the Greek letter λ.

Mathematically a wave is typically represented by the sine (or cosine) function. The box shows a MATHEMATICA plot of the function $\sin(x)$ for $0 \leq x \leq 2\pi$. The resulting graph has already been shown in Fig. 1.15.

MATHEMATICA
Plot[*Sin*[x], {$x, 0, 2Pi$}] shift-enter
as we have done in the chapter on geometry.

If we wish to plot the wave in such a way that we have exactly one full wavelength over a given unit distance, then we need to plot the function $\sin(\frac{2\pi x}{\lambda})$, because then we have a wave rising from 0 at $x = 0$, encompassing a maximum at $x = \frac{1}{4}$ followed by a minimum at $x = \frac{3}{4}$ and returning back to 0 again at $x = 1$. Thus $\sin(\frac{2\pi x}{\lambda})$ represents the mathematical formula for a wave as a function of distance. However, the wavy motion is also a function of time. At any given point x the wave moves up and down with the frequency ν. Therefore, in order to capture this time dependence, we need to insert the term $2\pi\nu t$ into the sine function to arrive at

$$\text{Wave} = A \sin\left(\frac{2\pi x}{\lambda} + 2\pi\nu t\right). \tag{2.39}$$

Here we have added a so-called amplitude factor A. This is because the sine function is at most equal to 1. If we wish the water wave to be 5 m high at the crest and 5 m deep in the valley, as it sometimes occurs in Hawaii and at other places, then we need to use $A = 5$ m to describe that wave. The amplitude A may also slowly vary with time and distance. For example, the amplitude of the Tsunami pressure wave may become smaller after a very long travel, simply because the energy of the wave is distributed over a larger area. We assume for most of our discussions that A is constant.

To understand how well Eq. (2.39) describes a wave, it is useful to perform calculations for fixed times and variable space coordinate x and conversely for fixed space coordinate and variable time coordinate t. This can be done nicely with any software that plots mathematical functions or with a little more effort even with a pocket calculator. Because one does not wish to repeatedly write the 2π factors when dealing with this type of equation, one defines the so-called angular frequency $\omega = 2\pi\nu$ and wave number $k = \frac{2\pi}{\lambda}$ to obtain

$$\text{Wave} = A \sin(kx + \omega t). \tag{2.40}$$

An important relation that is valid for all waves follows from the fact that the velocity v of the wave must be equal to the product of the wavelength and the frequency, simply because the wave makes ν full oscillations per unit time and each oscillation corresponds exactly to one wave length λ. Thus we have

$$v = \lambda\nu. \tag{2.41}$$

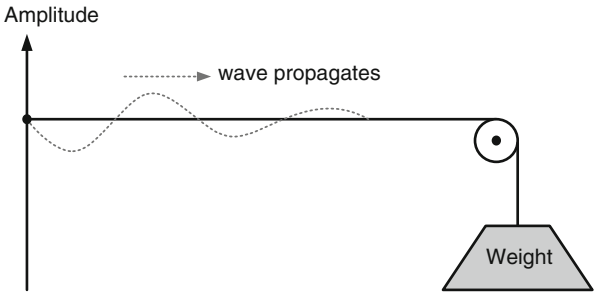


Fig. 2.12 A string that is fixed on one side and pulled by a weight on the other side can perform wavelike motions when excited. This happens, for example, to the string of a guitar. The wave that propagates from the point of excitation on the left to the other side of the string is described mathematically by Eq. (2.39). The amplitude of the wave may stay constant. Often, however, the amplitude decreases with traveling distance

This relation is also valid for light waves. Because we denote the velocity of light by c we have

$$c = \lambda \nu. \tag{2.42}$$

Strings and Standing Waves

Equation (2.39) describes a one-dimensional wave because it considers only the x-direction. One can experimentally create such a wave by using a long string that is held fixed at one end and pulled by a weight on the other end. Then one can hit the string at some point, exactly as one plays the string of a guitar, and a wave will start propagating away from that point. This effect is shown in Fig. 2.12.

Of course, this is exactly what happens when a guitar is played. However, guitar strings are very short, and the wave is immediately reflected from at the ends of each string. A reflected wave propagates in the negative x-direction if the original wave has propagated in the positive x-direction. The equation for the reflected wave is therefore

$$\text{Reflected wave} = -A \sin\left(-\frac{2\pi x}{\lambda} + 2\pi \nu t\right), \tag{2.43}$$

which is identical to the original wave except for the negative sign of the term that contains the space coordinate x. We also have changed the sign in front of the equation. This sign change comes from the fact that the reflection at the end of the string turns the amplitude of the wave around. The result that one obtains if one adds both waves together is the vibration pattern that the string of a guitar shows. Adding wave and reflected wave gives

$$\text{Standing wave} = \text{Wave} + \text{Reflected wave} = 2A \sin\left(\frac{2\pi x}{\lambda}\right) \cos(2\pi \nu t). \tag{2.44}$$

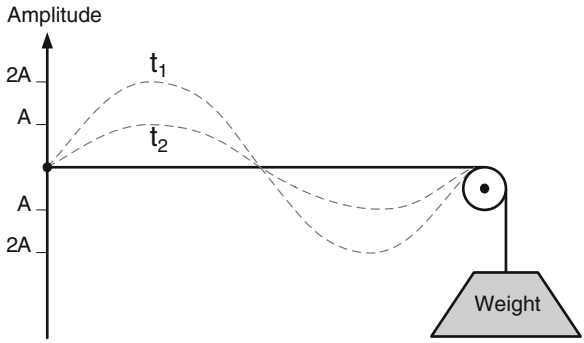


Fig. 2.13 A "standing" wave plotted for two different times t_1 and t_2 that are chosen such that $\cos(2\pi\nu t_1) = 1$ and $\cos(2\pi\nu t_2) = \frac{1}{2}$, respectively

This result can be obtained by using the so-called theorem of subtraction of the sine function, which tells us that $\sin(\alpha) - \sin(\beta) = 2\sin(\frac{\alpha-\beta}{2})\cos(\frac{\alpha+\beta}{2})$. The resulting wave is called a standing wave. The reason for this name is that the maxima (and also minima) of this wave do not move anymore. They stay always at the same place, meaning that the velocity of this resulting wave is 0. One can see this easily by plotting Eq. (2.44) for different times as done in Fig. 2.13.

The figure is plotted for the case that the wavelength of the exited wave equals the length of the string. It is also possible that the string is just half a wavelength long and the standing wave has just one maximum or minimum exactly in the middle of the string. This is actually the usual form of excitation when a guitar is excited by plucking it right in the middle of the string.

The frequency of the vibrating string is determined by the string tension (the weight stretching the string), the mass of the string, and its length L. Thus the musical tone that a string excites in air through its vibration also depends on these factors, because it is determined by the frequency of vibrations. The possible wavelength of a vibrating string is, of course, also determined by the length of the string, and we typically have $L = \frac{\lambda}{2}, \lambda, \frac{3\lambda}{2}, 2\lambda, \ldots$ It is known since Pythagoras that the tone of a string with length L is higher than that of a string with length $2L$; musically speaking it is an octave above. There are many interesting projects with vibrating strings, and the Internet is full of information about it. More complicated objects than strings, such as metal plates, drums, the body of a violin or a piano, vibrate in standing waves also. These standing waves are not just waves in one dimension but, in general, in three dimensions, and they have very complicated shapes. The complicated shapes are, in turn, the reason why such a world of sounds can be created by instruments like the guitar, the violin, a piano, or an organ made of pipes.

Standing waves play also a major role in modern physics and particularly in the area of quantum mechanics. One of the theories that is a contender to explain all phenomena that we know is the so-called string theory. Its name derives from the fact that elementary particles are described by using concepts of vibrating strings as we just discussed them. The interested reader is encouraged to surf the Internet and learn more. We also have added material in Sects. 5.3 and 5.4.

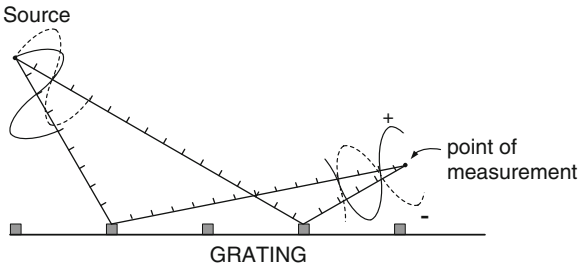


Fig. 2.14 A wave emitted from a source is interacting with a grating. A point of measurement of the wave intensity is also indicated. The *waves* are shown close to the source and also close to the point of measurement by *dashed lines*. The *full lines* with distance markers show the wave direction. The height of the two waves that are reflected from the grating is equal and opposite at the point of measurement and therefore the waves cancel each other resulting in zero total wave intensity at this point. If the wave would be a light wave, the point of measurement would be *dark*. This effect is called destructive interference

Diffraction of Waves

All the phenomena described above are important for waves. The one single phenomenon that is usually considered the hallmark of waves is diffraction. The word diffraction, or diffraction of waves, simply means that the wave is influenced in a specific and important way by some structure of a material that interacts with the wave. This structure could consist of wooden posts that influence water waves, thin wires that influence electromagnetic waves, or it could be just the atom layers of a diamond that influence light and make it glitter in beautiful colors.

A very important diffraction structure is the so-called diffraction grating. A diffraction grating is simply a number of lines engraved into some material. This can be straight lines such as metal lines painted on top of glass, or lines created as scratching the glass, or scratches in a thin metal layer on plastic, and so on. The lines may also be drawn perpendicular to each other, thus forming a two-dimensional crosswise pattern or even a three-dimensional pattern, such as a crystal lattice. The crystal lattice may either be artificially made, as described in our section on nanostructures (Sect. 3.4), or be just a natural crystal lattice, such as diamond. In other words a diffraction grating is just a geometrical structure with regular arrangements of lines of different atoms or even of missing atoms typically in two or three dimensions. Photons that hit the reflection grating interact then with a (large) number of atoms of these lines and, as a consequence of this interaction, are redirected and reflected. This redirection and reflection does not have any arbitrary direction. It occurs just in certain directions that depend on the spacing of the lines and on the wavelength of the waves that are incident on the grating. The reason for the different redirection or reflection into the various directions is that the waves that are redirected from many points of the grating may amplify each other in certain directions, while they may weaken or destroy each other in other directions. This latter effect is shown in Fig. 2.14.

In this figure we have included a source of waves and also two paths the wave can take toward a point of measurement. These waves could be, for example, microwaves, i.e., electromagnetic waves as they are used in microwave ovens. One path of the waves is 23 cm long while the other is only 21 cm long. The wavelength of the wave shown is chosen to be 4 cm. We can see from Fig. 2.14 that the two waves would end up with their electric fields pointing in opposite directions at the point of measurement. Therefore the two fields would cancel each other at that point, and no wave would be measured at this point. One calls this destructive interference of the waves. One actually does not need microwaves to see such an effect. One can also perform such an experiment with water waves. You can create waves with two fingers while sitting in the bath tub, and you will observe spots on the water surface that do not show any wave motion, because the reflections from the various walls of the tub cancel each other at these spots.

The calculation of interference effects proceeds as follows. We use Eq. (2.39) to describe the wave. Now we have two waves:

$$\text{Wave1} = A \sin\left(\frac{2\pi\, 23}{4} + 2\pi\nu t\right) \tag{2.45}$$

and

$$\text{Wave2} = A \sin\left(\frac{2\pi\, 21}{4} + 2\pi\nu t\right), \tag{2.46}$$

where A is just giving the highest amplitude of the wave that is obtained when the sine function equals one. We have in this equation also time t and frequency ν of the wave. However, the term containing time is for both waves the same and therefore not important for our discussion. We therefore put at first $t = 0$ and forget about this term. We can then calculate the sum of the two waves at the point of measurement and find

$$\text{Wave1} + \text{Wave2} = A \sin(11.5\pi) + A \sin(10.5\pi) = 0. \tag{2.47}$$

This result can be found by use of a pocket calculator or by MATHEMATICA as shown in the box.

MATHEMATICA
Sin[11.5 Pi] + Sin[10.5 Pi] shift return
Output = 0

You can also add the terms with any given time in the sine functions, and the result will still be 0. Actually calculators and also MATHEMATICA may give you some very small number very close to (but not quite) zero. This is because computers can often only calculate approximate numbers. At any rate, the two waves annihilate each other at the point of measurement.

Had we used a wave with different wavelength, say 0.5 cm instead of 1 cm, we would have obtained a different result. Then we would have had

$$\text{Wave1}+\text{Wave2} = A\,\sin(23\pi+2\pi\nu t)+A\,\sin(21\pi+2\pi\nu t) = 2A\,\sin(\pi+2\pi\nu t). \tag{2.48}$$

Now our result does depend on time as you can easily check by use of MATHEMATICA or a pocket calculator that features the sine function. The maximum result that you can obtain equals $2A$ and the minimum equals $-2A$. In other words, the amplitude of the wave has doubled because the two waves have added and helped each other. One calls this case constructive interference.

Note that the wavelength of wireless phones happens to be also around 1 cm. That means that interference effects with close-by walls may influence your cell phone receiving strength. The wavelength of visible light is much smaller around 600 nm. However, you can still perform the same experiment with light. Of course, then the spacing of the lines of the grating needs also be around 600 nm. The point of observation that we have chosen in the above example would give no light for the longer wavelength and brighter light for the shorter. Sunlight is actually a mixture of all light colors. Red light corresponds to the longer wavelength and blue light to the shorter. The point of measurement would therefore appear blue, and points of observation more to the right in Fig. 2.14 would be red. How can we make such an experiment? There is indeed such a fine grating in every household, because DVDs that are used to play movies contain the information for the pictures in the form of very closely spaced dots that are positioned on very closely spaced circles. Indeed if you hold a DVD toward sunlight you see all colors of the rainbow that are created because light with different color is reflected by the grating in different directions. If you have a laser pointer, then you can do an additional experiment: shine the laser pointer in a grazing angle onto the DVD and watch the reflections on a wall. If you had a mirror instead of the DVD, you would see exactly one point as the reflection of the laser pointer. However, now with the DVD, you can see several points often up to five or more and all in the same color because the photons of the laser have only this one color. You can calculate that there exist several points at which a grating causes the light to constructively interfere; this is done by just calculating the sum of two waves with one given wavelength λ for a sequence of observation points. You will see then constructive and destructive interference effects and therefore minima and maxima of light as you proceed more to the right in Fig. 2.14.

The Doppler Effect

We end this discussion of waves with a few remarks on another important effect that was discovered and investigated by the Austrian physicist Christian Doppler. If a person moves with the wave, for example, exactly as fast as one of the highest points of the wave, then to this person, the wave will appear to stand still. If the person moves in the same direction as the wave does, but a little slower, then the frequency

of the wave that person measures appears to be lower. The reason is simple: if you walk in the direction of the wave and count the highest or lowest points of the wave that pass you during a given time period, then you will count a smaller number as compared to the case when you stand still. The frequency of the wave that you measure is thus decreased because it is given by the number of highest or lowest points that you counted divided by the period of time during which you counted. If you walk fast enough in the direction that the wave propagates, you can even make it happen that no highest or lowest point of the wave passes you. Then, from your view, there does not seem to exist an up and down motion, and the frequency of the wave appears to be zero. Inversely, if you walk against the wave, then you count a larger number of wave maxima or minima while walking than when standing still. Thus the frequency of maxima or minima that you count is increased. That frequency difference that occurs in situations where we encounter waves and moving objects is called the Doppler effect.

The Doppler effect can be observed and used in a large variety of situations. For example, one can send electromagnetic waves toward a moving object and can conclude from the Doppler effect, in this case the change in the frequency of the reflected waves, how fast the object is moving. Such a measurement system, that sends out electromagnetic waves and measures reflected waves (including their frequency), is called radar. The frequency range of radar is typically in the GHz range. Doppler radar is used to determine the speed of approaching storms and the speed of winds in a hurricane. Doppler radar is also used by police to determine the speed of cars.

2.2 Electromagnetic Phenomena

The ancients knew already that one could rub a little block of amber, looking like a yellow stone, against a fur, against a towel, or against one's coat and then draw electrical sparks out of it. That little piece of material also tends to attract one's hair or even raise one's hair in all directions when it is dry. Also known to the ancients was that small metallic looking pieces of certain materials would attract each other. We now call these pieces magnets, and I remember how fascinated I was as a boy when I saw magnets for the first time. The excitement was even greater when I built my first battery-driven electromagnets. I put paper between the magnets and put my finger between them, and they still attracted each other while my fingers felt nothing. This was like magic to me. Astounding electrical phenomena were actually always known to mankind. The lightening of thunderstorms was appreciated as very special and, for example, attributed to the Greek god Zeus.

We know now that electricity and magnets are not magic, and we understand electricity and use it in our daily life. In fact the basic electrical particle that underlies the cause of all these phenomena, the electron, is all around us. It derives its name from the Greek word for amber which is "electron." Electrons are not only part of amber but of all materials that we know. Everything we touch, everything

we see, everything we smell, is somehow connected to electrons. How did that fact stay hidden from us for such a long time? How did mankind tame electricity for its purposes? These and other questions are the topics of this section.

Electricity was more difficult to figure out than gravity, the force that rules the motion of the planets, and has been explored in scientific detail by James Clerk Maxwell and Michael Faraday 200 years after Newton (also in England). A major reason for the difficulties to understand electricity is the magnitude of the force related to it. This force is actually much much larger than the gravitational forces between two massive bodies such as earth and sun that is described by Newton's equation (Eq. (2.37)). In contrast to the forces of gravity, however, there exist two opposite types of forces between electrical particles that are the sources of these forces. Depending on which kind of electrical particles one deals with, one obtains attractive or repulsive forces. One distinguishes electrical particles that are "positively charged," meaning that they are the sources of one type of electrical force, and particles that are "negatively charged" and are the sources of the opposite type of force. Opposite means here that the forces point in opposite directions. Positive and negative charges attract each other, while positive charges repel other positive charges and negative charges repel other negative charges. It now so happens that the most basic atom, the hydrogen atom, is composed of one particle with negative charge, that is the electron that we mentioned above, and one particle with positive charge, that is called the proton. This means that the sum of charges that are contained in the hydrogen atom is zero, and this fact is also true for all other atoms. Atoms are extremely small and cannot easily be decomposed into their parts, because the positive and negative charges attract each other and stick closely together. This is the reason why electrical charges and their presence in all atoms were not recognized earlier. In fact, Maxwell himself still thought that no electrical charges would be present in metals! This is a very incorrect notion. Metal wires carry the electrical currents (flowing charge) that supply whole cities with power.

Why is the hydrogen atom of such basic importance for the understanding of electricity? Hydrogen atoms are the most abundant type of atoms in the universe. There exists a silly joke that only stupidity is more abundant in the universe than hydrogen. Our sun and billions and billions of similar stars in each of the trillion of galaxies are mostly composed of hydrogen. We will learn in Chap. 5, that deals with Einstein's theory of relativity, that a very special mechanism involving hydrogen makes the sun and the stars shine. Most of the other atoms that we know have probably been generated from hydrogen during the aging of the stars and during explosions of stars (look up "supernova" on the Internet). Because hydrogen is composed exactly of one positive and one negative electric charge, we can therefore deduce that the number of positive and negative charges in the universe is about the same. Charge is, in addition, conserved in all physical processes, meaning that the number of positive and negative charges stays the same. This is the law of charge conservation, which is held almost at the same level of importance as energy conservation is.

Positive and negative charges attract each other with a very significant force. This is why the electron and the proton of hydrogen are very close to each other,

and one needs a considerable energy to separate them over larger distances. This is also the reason why atoms were thought to be indivisible for a long time. Typically the separation of protons and electrons in hydrogen atoms is less than 10^{-8} cm or equivalently 10^{-10} m. Just as an aside, because one speaks a lot about nanoscience and nanometers, 1 nm is 10^{-9} m, and therefore the separation of electrons and protons in the hydrogen atom is less than 0.1 nm. We will learn more about atoms and their constituent charged particles in following sections. The electron and the proton are the origin of opposite electrical forces and, because they are opposite, these forces originating from close-by electrons and protons (as they are in atoms) cancel each other at distances far away from the atoms. It is for this reason that under normal circumstances we do not notice electrical forces, in spite of the fact that all matter that we know and that surrounds us on earth contains electrons and protons.

Only when we separate electrons and protons can we see the enormity of the electrical force. This separation takes energy. We will learn later how to calculate this energy. For now we note only typical values. It takes the energy of about $2 \cdot 10^{-18}$ J to separate the electron and proton of the hydrogen atom. A small glass of water contains about 10^{25} hydrogen atoms. To separate then all the protons from the electrons and put them into two different containers will take the energy of about $2 \cdot 10^{7}$ J. That energy would heat a kitchen stove for more than an hour. It is therefore understandable that (a) we normally do not see or feel the electrical forces, because the positive and negative charges are not being separated and (b) if the charges are separated, then they may be useful to heat our kitchen's stove. In fact, electrical forces are now all pervasive in our daily life and do for us a lot more than just heating a stove. The question for such applications is therefore, how do we separate positive and negative charges and how do we generate electrical forces?

2.2.1 Galvanic Elements: The Lithium Battery

Historically, an important generator of electricity was the so-called galvanic element, named after its discoverer Luigi Galvani from Italy. Batteries that we use in our daily life for cars or laptops, are based on the principle of galvanic elements: the generation of electrical power and currents by chemical processes. A major difficulty in the understanding of galvanic elements arises from the fact that one needs to understand some chemistry to start with. One needs to know a bit about the nature of atoms and the nature of combinations of two or more atoms to form so-called "molecules" and "chemical compounds." However, an important part of chemistry itself needs to be explained by describing experiments that involve galvanic elements. The methods of explanation of all of these facts involve, therefore, a "circle" that one needs to master by first accepting some of the chemical terms without detailed explanation in order to understand the basic electric phenomena, and then reading up on the chemistry later.

We know that the most important atom, the hydrogen atom, consists of one negatively charged electron and a positively charged proton. We have mentioned this above, without emphasizing the surprising fact that an atom is *not* indivisible, as the ancient scientist Democrit thought, but can be decomposed into electrons and protons. It turned out that this fact was a most important discovery. All the materials that we know are composed of atoms, and all of these atoms contain electrons and protons. The major chemical differences of atoms arise simply from the fact that they are composed of different numbers of electrons and protons. The helium atom has two electrons and two protons, and the next atom is lithium with three electrons and three protons. For reasons that we will discuss in Sect. 2.4 on chemistry and quantum mechanics, one of the three lithium electrons is rather easy to remove from the atom. It takes only $\frac{1}{3}$ of the energy to separate one lithium electron from the lithium atom as compared to the energy that is necessary to remove the electron from hydrogen. The separation energy for a helium electron, on the other hand, is about twice that of hydrogen. Lithium is therefore special and is at room temperature a solid metal while both hydrogen and helium are gases. The small atom size and the metallic properties make lithium useful to produce galvanic elements that we nowadays just call batteries. A lithium battery most probably powers your laptop and, in the future, may possibly power your car.

Originally, batteries consisted of metals embedded in a liquid in which a salt of the metal was dissolved. Salts are, of course, very well known to everyone. The salt used for cooking is sodium chloride, consisting of a sodium (Na) and a chlorine (Cl) atom, and is denoted in chemistry by the symbol NaCl. Sodium is a metal just as lithium is. In liquids, the sodium and the chlorine atoms dissolve and separate. However, the atoms in the solution are now charged, because one sodium electron can easily be removed (just as one electron of the lithium can be). The sodium becomes then positively charged, and the chlorine takes that electron and becomes negatively charged. This preference of the electron to move from the sodium to the chlorine has some deep chemical reason that we will discuss in detail later. Thus, if you drop salt into water, the water will then contain positively charged sodium atoms denoted as Na^+ and negative chlorine atoms denoted by Cl^-. One calls such a liquid (water plus salt) an electrolyte, because it contains elements of electricity and conducts electrical currents. The dissolved and charged atoms Na^+ and Cl^- are called ions. Ions are simply atoms carrying charge. The same effects happen if you dissolve a lithium salt instead of NaCl, or any other salt.

The principle of the workings of a lithium battery (and a large number of batteries based on different materials) is shown in Fig. 2.15. Lithium atoms that are contained in some chemical compound (Ch1) on the left side of the battery dissolve into the electrolyte leaving an electron behind at that left contact side that is labeled in car batteries as “black.” Black indicates that a negative charge consisting of electrons is accumulating at that contact. The dissolved lithium atoms have lost these electrons and become, therefore, positively charged ions. The electrons propagate via a wire to the lightbulb (car, laptop, or whatever is powered by the battery), and the positively charged ions propagate within the battery toward the right contact. At the right contact, that is labeled in car batteries as “red,” the electrons that have moved

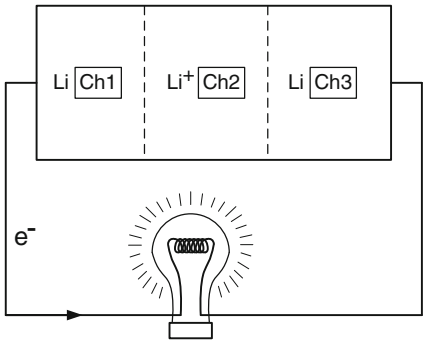


Fig. 2.15 Principle of a lithium battery: The *rectangle* indicates the battery that has two electrical contacts. The *left* contact supplies negative electrons indicated by e^- and is often labeled by the color *black*. The other contact (to the *right* in our figure) is positive and usually labeled by the color *red*. At the *black* side, lithium in the form of a chemical compound indicated by (Ch1) loses one electron and proceeds as positively charged lithium Li^+ through another chemical (Ch2). (Ch2) permits Li^+ to pass but rejects electrons. Li^+ arrives at the red (the positive) side. The negative electron, on the other hand, proceeds via the outer circuit and the lightbulb. It finally arrives at the right side and neutralizes the positively charged Li^+ that together with (Ch3) forms the compound Li(Ch3) as indicated. The flow of many electrons that is generated that way can be used to power a lightbulb or a personal computer and the like

through the lightbulb (or other devices) and the positive ions that moved inside the battery reunite to form again complete lithium atoms with help of chemical (Ch3). The chemical processes that are involved in all of this electron and ion generation create thus an electrical current in the wire that supplies the necessary power to drive our chosen equipment. In this way chemical energy is turned into electrical energy, and the chemicals need to be carefully chosen to achieve this.

For safety reasons one can not use pure lithium-metal contacts in batteries. Lithium metal would burn in air explosively. This is another important reason why battery producers use chemical compounds of lithium, meaning they use lithium atoms connected to a number of other atoms. In our Fig. 2.15 the chemical compounds are just denoted by (Ch1), (Ch2), and (Ch3), because we are here mostly interested in the electrical current generation and not in the chemistry of these compounds. The compounds, which in principle can be solids, liquids, or even gases, determine in the final analysis the performance of the battery, how much power the battery can deliver, and how often it can be discharged and recharged. The design of battery materials is a very important area of STEM, and many chemical engineers attempt currently to produce the most competitive batteries.

The amount of electrical current that can be drawn from a battery depends on the number of lithium atoms that can be dissolved. Typically the design is such that a fraction of the total available lithium dissolves per hour, so that the battery can work for several hours. To give an example, we assume that the total available number of lithium atoms at the left contact is given by Avogadro's number of $6.022 \cdot 10^{23}$ atoms which corresponds to a few grams of lithium (see Sect. 2.3.2). Assume further that

a small percentage of these lithium ions dissolve into the electrolyte say 4 % per hour. Then about $2.4 \cdot 10^{22}$ positive lithium ions propagate per hour toward the red (positive) side of the battery. At the same time the same number of electrons flows through the lightbulb or our laptop. This corresponds to about $6.69 \cdot 10^{18}$ electrons and ions per second. These are all very reasonable numbers, and this is what actually happens when you power a lightbulb or your laptop with a lithium battery. However, you are usually given the electrical current in units of amperes, not by the numbers of ions and electrons per second. We will return to this important unit of ampere below.

The chemical equation that describes the discharge of a lithium battery at the left (black, negative) side is

$$\mathrm{Li(Ch1)} \leftrightarrow (1-x)\mathrm{Li(Ch1)} + x\mathrm{Li}^{+}\mathrm{(Ch2)} + xe^{-}. \tag{2.49}$$

Here, the double-sided arrow means that we can read the equation in both directions. If we use the battery to power lightbulb or laptop and discharge it, you must read the equation in the direction toward the right $\rightarrow$. Lithium dissolves in the form of x % of positively charged ions into the electrolyte (represented by the term $x\mathrm{Li}^{+}\mathrm{(Ch2)}$) and leaves xe^{-} electrons at the contact that supplies wires toward the lightbulb or laptop. The symbol e^{-} indicates the negative charge of one electron. The number of the electrons that are generated is, of course, determined by the percentage x that dissolves. For example, if 4 % are dissolved, then $1 - x$ represents 96 % of lithium that is left over and is represented by the term $(1 - x)\mathrm{Li(Ch1)}$.

At the right side of the battery (red, positive), the following equation describes what happens:

$$(1-x)\mathrm{Li(Ch3)} + x\mathrm{Li}^{+} + xe^{-} \leftrightarrow \mathrm{Li(Ch3)}. \tag{2.50}$$

The positive lithium ions, that originate from the left side while the battery is discharged, are neutralized by the electrons that have been propagating through the lightbulb or laptop and are now returned at the right side. There, the resulting lithium atoms form the compound Li(Ch3).

Up to now we have only considered the process of discharging the battery that produces electrical power. The inverse process is also possible. We use electrical power from a so-called recharging equipment to restore the battery to its original charged state. The chemical equations for the recharging are also given by Eqs. (2.49) and (2.50), but now the equations need to be read from right to left $\leftarrow$. On the left side of the battery (with Eq. (2.49) being the relevant equation), we can supply negative electricity in the form of electrons from the recharging equipment. This negative electricity attracts and subsequently neutralizes positively charged lithium ions that thus are returned as lithium atoms and restore the original Li(CH1). On the right side of the battery, electrons are extracted by the positive recharging equipment, resulting in positively charged lithium that is dissolved and propagates now to the left while the extracted electrons are transferred by the recharging equipment toward the left contact.

Two important consequences follow from the above description of the lithium battery that should be remembered by everyone even if the details of the process

are forgotten: (a) Batteries have a positive contact labeled as "red" contact or by a "+" and a negative contact labeled as "black" or by a minus "−." The black is connected in a car to all of the metal of the car and often called the "ground" contact. (b) The recharging equipment also usually has a red or positive and a black or negative contact. When recharging the battery you must connect red to red (positive to positive) and black to black (negative to negative)!

What we have learned up to now is that one can construct a chemical machinery that results in a flow of charged particles. There is energy involved in that flow. In one direction of the flow we can operate, for example, a lightbulb and use the energy. In the other direction of flow, we need to supply energy to the system by using the recharging equipment that we usually plug into a power outlet of our home. What are the energies that are involved in this flow of charge? The chemical reactions in the galvanic element give rise to an energy difference between the left and the right contacts. Each electron that flows from the left to the right gains that chemically generated energy and can, in principle, transfer it to the equipment that we have attached to the galvanic element. That chemical energy depends on the chemicals that one uses for the galvanic element and is typically $2.4–4.8 \cdot 10^{-19}$ J for each electron. This is a very small decimal number, and one can understand that the average person will not know what to do with such a number or what it means. For this and historical reasons, one uses different units for the energy that is supplied to electrons. Instead of Joules, one uses electron volts or eVs. The typical energy that each electron gains in a galvanic element is 1.5–3 eV, and one eV corresponds to $1.602 \cdot 10^{-19}$ J. Electrical engineers are usually not talking about energies of electrons but just about the "voltage." A battery supplies 1.5 V (V for "volt") means that each electron that propagates from the negative to the positive side gains 1.5 eV. This energy gain gives us some idea that energy and power can be supplied by batteries. We know that it is not dangerous to touch a battery that supplies 1.5 V. It just stings a little if we put two wires from a battery to our tongue. However, a power outlet that supplies typically 110 V is not to be touched without great danger. The thousands of volts of a high power line are even more dangerous, and accidents related to these power lines have killed people.

As you could see, the movement of ions and electrons in batteries is not so easy to understand. The chemistry of batteries is an even more complicated subject, because it involves a lot of chemicals and a lot of possibilities of interactions between them. What one wishes to have is a battery that can store the largest possible energy and does not degrade with frequent recharging. For example, a battery that powers a car would need to supply enough power to drive about 100 km or even miles. It should at least be possible to charge it 2,000 times because then the car could be driven 200,000 km (or better 200,000 miles) before the battery needs to be replaced, which is very expensive. Ideally, the battery should be, of course, as inexpensive as possible and mass producible. Because all these requirements are difficult to meet, more research and development by STEM experts is needed in this area. Recent developments of new types of batteries that involve so-called nanostructures do appear very promising.

2.2.2 Basic Concepts Related to Electricity

The number of electrons that flow is also inconveniently large and is therefore not used in the language of engineers and in daily life. One uses the unit of A for "Amperes" to indicate how much charge is flowing in one second from one electrical contact to the other. The unit of charge therefore becomes *As* standing for "Ampere seconds", simply because we have to multiply the charge flow during one second by the number of seconds, in order to obtain the total charge that is transferred from one contact to the other. Typically we have a flow of 1–10 A through a lightbulb or a laptop, when we power them with a battery. The charge of a single electron is $-1.602 \cdot 10^{-19}$ A s (Ampere seconds). (One often denotes the positive value of this charge by the letter e and speaks of the elementary charge.) To obtain the typical 1–10 A, we need then, as we know, a large number of electrons. To obtain 1 Ampere second of charge, or as one says a "current" flow of 1 Ampere, one needs to multiply the charge of one electron by the number $6.24 \cdot 10^{18}$ which is then equal to the number of electrons that actually flow. As with all units, it takes a while to get used to them. Units are (or at least should be) chosen to be convenient. However, the "convenience" depends on what we are actually doing with the units, and with electrical energy we are doing many different things. The operation of lightbulbs was very important in the history of electricity. Nowadays we have an enormous number of applications that are powered by electricity. This fact becomes very noticeable when we have a power failure.

What is the definition of electrical power? It is simply the energy that can be and is supplied per second. The electrical energy E_{elec} that is supplied according to our discussion above is given by the energy that is supplied per charged particle (electron) multiplied by the number charges; thus

$$E_{\text{elec}} = \text{V A s}. \tag{2.51}$$

In words, the electrical energy that is generated in every second is given by the number of volts that the galvanic element or power-outlet supplies, multiplied by the charge that flows which is measured in ampere seconds A s. The power to do things depends, of course, on the energy that we have available per second that means the energy that we have obtained divided by the number of seconds during which it was delivered. Thus the electrical power is given by:

$$\text{Power} = \text{V A} = \text{W}. \tag{2.52}$$

In words, the power is measured in volt amperes (volts times amperes) which is also called "watts." All these names are derived from the scientists and engineers that developed the knowledge and use of electricity. If you have a given voltage, as, for example, the voltage of 110 V of a standard US power outlet, and if you have a given power that some appliance takes, say 1,100 W, then you can calculate the electrical current (charge flow) to be 10 A ($= \frac{1{,}100\,\text{W}}{110\,\text{V}}$). Sometimes one needs to know all these values to buy the fitting appliance and the wires with the appropriate thickness.

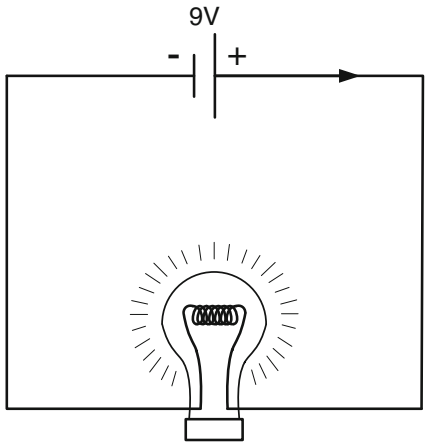


Fig. 2.16 Electric circuit with lightbulb (or other light emitter) driven by a galvanic element (battery) that is represented by the symbol on top of the figure. The + sign indicates the positive electrode of the battery and the − sign the negative. By convention, the *longer vertical line* | always indicates the positive side. The voltage of the battery is also shown on top and is, in this case, 9 V

Lets look, therefore, at some typical values of the power that we need to operate frequently used devices or instruments. The lightbulb of a typical lamp on our desk uses the power of about 20–40 W or equivalently 20–40 V A (volt ampere). A laptop uses about the same power and corresponding energy. The energy is obtained by multiplying watts by the number of seconds used. This also gives you the energy in Joules. If we wish to light up a whole room, we need more power and buy a 100–200 W lightbulb. Actually, there are different types of lightbulbs. The ordinary lightbulbs that have a glowing wire need a lot of energy, because much of the total energy consumed is turned into heat, and only a small fraction of the energy is turned into visible light. New types of lights, as, for example, light-emitting diodes (LEDs), do not produce much heat and need only much lower power. The author anticipates that light involving heated wires will fade out of history. It is simply too wasteful by generating more heat than visible light. To heat the home takes much energy and if we are doing it electrically we need about 10,000 W per home, again depending on the heating equipment. A TV will use around 300 W and a typical personal computer (PC) 100 W. Knowledge of the power consumption is important for decisions that we have to make in daily life, and it is good to remember volts, amperes, and watts and their meaning.

We usually do not need to know the details of what happens with atoms and chemistry if we deal with electrical equipment. As we could see from the above discussion, we need just the volts that the battery or any other power source supplies and some way to calculate the current, the amperes, in order to determine what we need to know about the electric energy and power. The complicated description of a battery in Fig. 2.15, that also describes some of the chemistry of the galvanic element, can thus be replaced by use of simpler symbols that are sufficient for the purposes of electrical engineering. The "electrical" version of Fig. 2.15 is shown in Fig. 2.16. To really use such electrical circuits just as the electrical engineers do, we need to solidify some of the concepts introduced above and also to introduce new concepts.

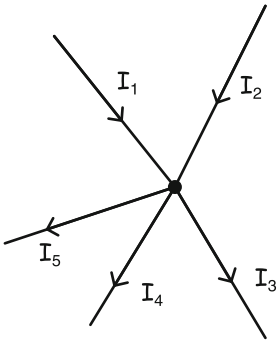


Fig. 2.17 Electric currents flowing toward and away from a node. A node is a point at which a number of current carrying wires are connected. Currents flowing toward the node are defined as being positive, while currents flowing away from a given node are counted as negative

We first discuss the concept of the electrical current. We know from Fig. 2.15 that we have essentially two differently charged particles that are the root cause of the electrical current. Within the battery we have positively charged lithium ions that are moving, and in the metal wire the negatively charged electrons move. If one investigates the chemical details of the galvanic elements further, one notices that there are also negative ions in the chemical Ch2. This is absolutely necessary because charges exert such a great force on each other that one cannot separate positive and negative charges, and big objects such as a battery need to contain therefore just about an equal amount of both charge types. This is also true for metal wires. Wires contain also positive charges, in the nuclei of their atoms (protons), which keep the wire overall neutral. All that matters for the electrical currents, power, and energy is the movement of negative and positive charges relative to each other. This is important to remember. Therefore, if we talk about electrical circuits, we usually do not need to know the actual movement of all the charges. All we need to know is how they are moving relative to each other. It is the historical convention (originating from times when nobody knew that electrons and protons existed) to assume that electrical currents are carried by the positive charges, while the negative charges are just there to keep the objects electrically neutral. Figure 2.16 shows, therefore, the arrow that represents the direction of the electrical current pointing from plus (+) to minus (−).

It is important to know that electrical currents behave in several ways completely analogous to streams of liquids. We have already discussed this analogy in Chap. 1.2 in connection with Fig. 1.6. There we have stated that the total current I_1 of water of a pipe, that branches out into two pipes (with currents I_2 and I_3, respectively), is conserved, meaning that $I_1 = I_2 + I_3$. This is a very plausible rule, that means simply that no water is lost. It also applies for electric currents and means then that no charge is lost either. Gustav Kirchhoff formulated this law in a more general way that is very valuable for electrical circuits. Figure 2.17 shows a number of wires that all are connected in one point, a "node." We define now all currents that are flowing toward the node as positive currents and all of those that are flowing away from the node as negative independent of what charges are moving. Kirchhoff's rule states that, as in the case of liquids, the total current flowing toward the node equals the

total current flowing away from the node. With the sign convention for the currents that we just discussed, this means that the sum over all currents at a node is zero. If we have N such currents then Kirchhoff's rule is:

$$\sum_{n=1}^{N} I_n = 0. \tag{2.53}$$

Resistors

The concept of electrical resistance of wires has also its clear analogies to the transport of fluids in pipes. A very thin wire has typically a higher resistance than a very thick wire made of the same material, just as a very thin pipe or tube cannot carry as much water per second than a very thick pipe can. Of course, the quantity of water that is actually flowing through a pipe depends also on the pressure difference between the two ends of the pipe. For electrical phenomena, the amount of current flow through a wire depends on the voltage difference between the two ends of the wire. Usually the stream of water that flows is directly proportional to the pressure difference, meaning that with twice the pressure difference we have twice the amount of water flowing. The same is usually true for electricity. With twice the voltage difference, the electrical current in a wire doubles. We say here "usually" because there exist many exceptions. For example, if the pressure becomes too large, the pipe explodes. Similarly, wires can burn out if the voltage becomes too large. However, even before that happens, a wire normally heats up when large currents flow and the resistance of the wire changes with the heat. Nevertheless, if we do not need to worry about such exceptions, the electrical current is proportional to the voltage. One then can define the resistance R of a piece of wire or any equipment by

$$R = \frac{V}{I} \tag{2.54}$$

which can also be written to express the proportionality of current to voltage:

$$I = \frac{V}{R}. \tag{2.55}$$

This equation is commonly known as Ohm's law, because it was Georg Ohm who noted in 1827 the proportionality of current I to voltage V. The unit of resistance is therefore the Ohm, denoted by the Greek letter Ω, and we have $1\Omega = \frac{1\,\mathrm{V}}{1\,\mathrm{A}}$.

If one is interested in electrical phenomena, it is important to remember a few typical values of resistance. A 100 W lightbulb has a typical resistance of 100 Ω. One meter of thin wire has about the same resistance, while a thicker wire may just have about 1 Ω or less. The resistance will also depend on the material that the wire is made of. Wires are made out of metals that are usually great conductors of electricity, meaning they are having a low resistance. Copper is a very good conductor and therefore often used for the electrical wiring in the household.

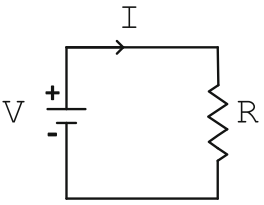


Fig. 2.18 Circuit with one galvanic element (battery) with voltage V and one resistor with resistance R. The current I flowing through the resistor is given by $I = \frac{V}{R}$

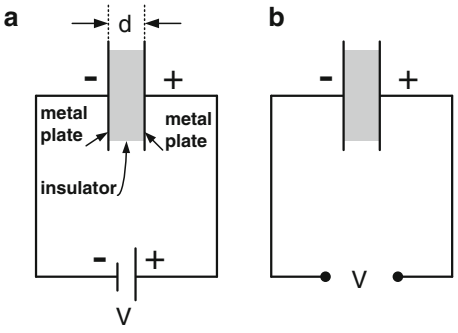


Fig. 2.19 (**a**) Two metallic plates, mounted parallel to each other and separated by an insulating (air or other insulator) region of thickness d, are called a capacitor. The capacitor is "charged" by the battery that supplies electrons to the left side and takes electrons away from the right until the voltage on the capacitor equals that of the battery. (**b**) The battery from (a) is removed. The voltage V can now be measured on the plates of the capacitor

A simple electrical circuit with one resistor is shown in Fig. 2.18. This circuit points toward another very useful rule which says, that in any closed loop such as given in Fig. 2.18, the voltage of the battery is equal to the product IR of the current and the resistance through which the current flows. This rule can be extended to more than one galvanic element and more than one resistor as discussed in Sect. 5.5.1.

Capacitors

A capacitor consists, in principle, out of two metal plates that are mounted parallel to each other at a distance d, as shown in Fig. 2.19. Also shown in Fig. 2.19 part (a) is a battery that forces the flow of charges onto the capacitor, until the voltage caused by the charges on the capacitor plates equals that of the battery. In part (b) of the figure, the battery is removed, leaving the capacitor charged and exhibiting the voltage V and a positive charge $+Q$ on one plate as well as an equal negative charge $-Q$ on the other plate. Both charges are, of course, measured in Ampere seconds A s. The capacitance C, which measures the power of the capacitor to store charge, is defined as:

$$C = \frac{Q}{V}. \tag{2.56}$$

The unit of capacitance is the Farad, named after Michael Faraday. One Farad is the capacitance that shows a voltage of 1 V on the plates, whenever the charge Q on the plates equals one Ampere second.

Capacitors have many applications in electrical circuits. For one, they can be used instead of batteries when it is important to recharge quickly. Capacitors can be charged almost instantly, because there is no chemical process connected with their charging. In the past, the standard capacitors could not store much charge. Recently, however, with the use of methods from nanoscience, it has been possible to create structures that have a very large capacitance. These structures are called super-capacitors. The secret is to be able to produce structures with very large area in a very small volume. Twice the area can store twice the charge. Small volumes are needed to reduce the distance between the plates which increases capacitance. Small volumes also make the capacitor useful for smaller equipment, to power toy airplanes and the like. Capacitors are needed otherwise in virtually any circuit application, ranging from the storage of charge in computer memories to the creation of electrical oscillations in wireless communications.

Currents are flowing to and from a capacitor only during charging or discharging. If one applies a constant voltage, then the capacitor charges up and the current stops flowing. If, on the other hand, one applies an alternating voltage, i.e., a voltage that varies with time from positive to negative, then a current that charges and discharges the capacitor is flowing continuously. Capacitors are, therefore, useful in electronic applications that let alternating currents (ac currents) flow and block constant or so-called direct currents (dc currents).

Several other applications (e.g., electronic memories that use capacitors) will be discussed in later sections. Here we finish with a neat recent discovery about capacitors. One can produce now very tiny structures, nanostructures, and therefore also very tiny capacitors. Remember that one nanometer is 10^{-9} m. We will discuss the production of such small structures in Sects. 3.2.2 and 3.4. It is possible to produce capacitors with a capacitance of Atto-Farad meaning 10^{-18} F. If we transfer exactly one charge (one electron) from one plate of the capacitor to the other, then the voltage that is necessary to do that is according to Eq. (2.56) given by

$$V = \frac{Q}{C} = \frac{1.60210^{-19}\,\mathrm{A\,s}}{10^{-18}\,\frac{\mathrm{A\,s}}{\mathrm{V}}} = 0.1602\,\mathrm{V}. \tag{2.57}$$

This means that if we apply less than 0.1602 V, no electron can be transferred from one capacitor plate to the other, and therefore no lasting currents can flow, because electrons *cannot* be divided into smaller entities with smaller charges. However, above that voltage, one electron can be transferred. Using higher voltages one can transfer one electron after the other like putting pearls on a chain. The neat experiments connected to this effect do not only tell us much about the existence of single electrons but also let us measure the effects of single electrons in molecules. This may lead to new applications in electronics and chemistry.

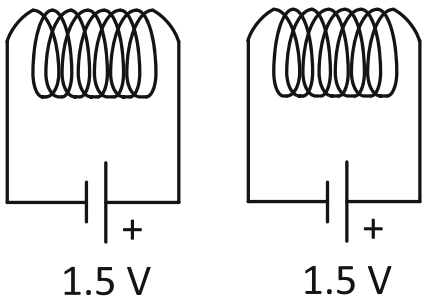


Fig. 2.20 Two electromagnets that can be created by wrapping insulated copper wires around a pencil or other round object. The electromagnets will attract each other when connected to batteries as shown

Electromagnets, Inductors

The following experiment is easy to do. Take some thin (about 0.01 cm) insulated copper wire and wind it (always in the same direction) around a pencil in about 50 layers with 10 windings each. This is schematically shown (one layer only) in Fig. 2.20.

Take away half a centimeter of the insulation at each end of the wire. All you need now is a battery, such as AA in the USA that provides about 1.5 V. Connect the ends of the wires to the ends of the battery, and you have produced an electromagnet that can pick up little pieces of iron or steel, such as paper clips. One also can influence the needle of a magnetic compass with this electromagnet, and, if you build a second identical magnet, you can do even more interesting experiments. For example, these two electromagnets will attract each other if you put the tail end of one to the front end of the other as also shown in Fig. 2.20.

I still remember the day that I made my first electromagnets. As soon as I noticed the force of attraction between the magnets, I was astounded. There was nothing in between them, yet there was a considerable force of attraction. Even more amazing, when I exchanged the battery connections of one electromagnet, then they repelled each other. Nothing on the outside of the wire-winded magnets gave away that something so special would happen. I put paper in between the magnets and nothing changed, the force went right through it. It appeared that one could exert some influence with one magnet over the other in spite of putting obstacles in between them. Did the magnets influence each other over a distance with nothing going on in between them? Or was something going on in between the magnets that was just not influenced by the paper? The following experiment provides a clue. Put the bottom of a steel pan in between the magnets. Then, if we turn off the power to the left magnet (disconnect the battery) the right magnet is still attracted to the steel pan but about the same way with and without the left magnet (be it switched on or off). From this it is clear that it does matter what is in between the magnets. Faraday was one of the great inventors and discoverers of electromagnetic phenomena, and he thought about it very deeply. A famous insight of Faraday is shown in Fig. 2.21.

Faraday surrounded the coil representing the electromagnet by iron dust. The little iron pieces start to form lines indicating the magnetic force everywhere.

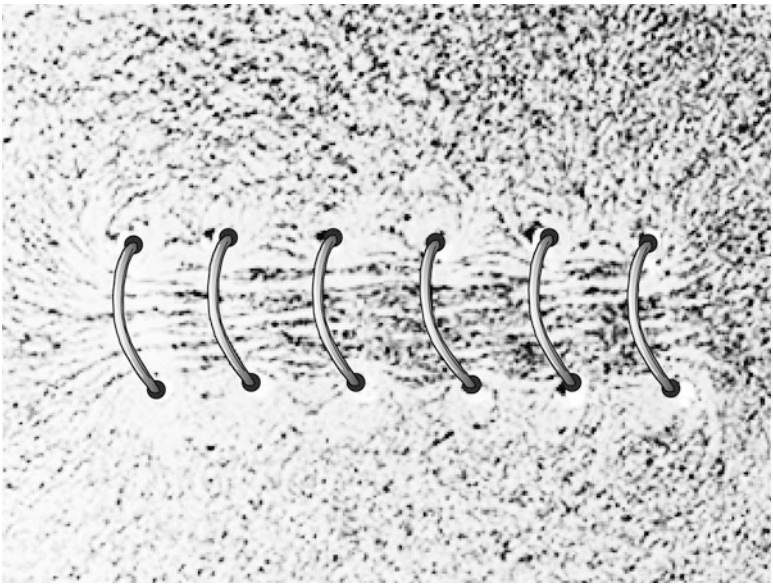


Fig. 2.21 Electromagnet surrounded by iron dust. Notice the formation of lines of dust corns that indicate the presence of the magnetic force

Faraday deduced from this that an electromagnetic field H is surrounding the coil at every point in space. He interpreted the interaction of one coil with the other as a consequence of the immediate field and not as an influence at a distance.

The question whether or not influences at a distance exist, or whether all influences are due to force fields that act only locally, is a very interesting one. Historically, gravity was conceived by Newton as influence at a distance. Faraday and James Clerk Maxwell, however, showed that influences at a distance were not necessary for their theory of electromagnetism, and Einstein was convinced that both gravity and electromagnetism should be understood without action at a distance. The discussion of all of these possibilities are still ongoing, and today's physics is divided on the point of influences at a distance. A number of scientists working in the area of quantum mechanics believe that influences at a distance are a possibility or even a necessity. This author believes that Einstein was correct, and nature can indeed be understood without any influences at a distance. I developed this conviction in spite of the fact that the action of magnets on each other appeared to me almost spooky in my childhood.

The experiments of Faraday (and many different experiments in the subsequent centuries) showed that the magnetic field H is created by the coils as a consequence of the electric currents that are flowing in the wires. A further interesting experiment can be performed as follows. We can superimpose the two coils on top of each other by taking two equally long wires and wrapping them parallel to each other around a pencil. Then, if we apply the batteries to both coils, we have a magnet that is twice as strong as that of any single coil. However, if we apply one of the batteries in the

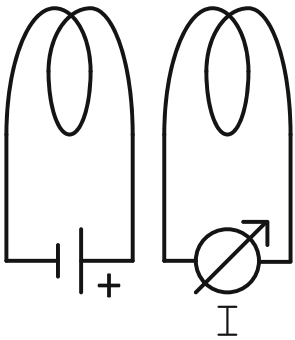


Fig. 2.22 Two coils that are very close to each other or even wrapped in parallel. One coil is connected to a battery, the other to an instrument that can measure current I or voltage V

opposite way as compared to the other, then we obtain no magnetic field. This is understandable because then the electric currents in the two coils flow in opposite direction and cancel each other out; they interfere with each other with the result of zero magnetic field.

The situation becomes even more interesting if we perform experiments that are time dependent or in other words, if we have a magnetic field $H = H(t)$ that depends on time t. This experiment requires two coils close to each other and is illustrated in Fig. 2.22. One coil can be connected to a battery, the other is connected to an instrument that can measure a current I. Such instruments are available in hardware stores and can usually measure not only currents but also voltages (as well as the resistance of a wire). After the battery in the circuit of the first coil is switched on, a current starts to flow and a magnetic field starts to build up. This buildup of the magnetic field takes energy and therefore does not occur instantaneously. It is as if the coil resists the magnetic field buildup. This "resisting" depends on how many loops such a coil has and on other circumstance is called the "inductance" L of the coil. The inductance is an important circuit element. Thus an inductive resistance is only present when there is a change in current and magnetic field.

During the time when the magnetic field changes in the first coil, a very important effect occurs also in the second coil. The instrument of the second coil indicates a current! One says that the current is "induced" in the second coil by the changing magnetic field of the first coil. Another way of expressing this fact is to say that the current is generated in the second coil by "induction." The inductance of coils in electrical circuits comes thus into play only when the electrical currents and corresponding magnetic fields change, i.e., when we deal with alternating voltages and currents. This is similar to the situation with capacitors that we discussed above. Note that there is no connection between the wires of the two coils, yet changes of the electrical currents of one coil cause electrical currents to flow in the other. The electrical current flows only if a force is present that accelerates negative (electrons) or positive (ions, protons) charges. The force per charge is called the electric field F. Thus a change of the magnetic field generated in the first coil "induces" an electric field and consequently an electric current in the second coil. The coils can even be spatially separated and the experiment still works, as long as the magnetic field that is shared by the coils is large enough.

It is important to remember the following fact. If one measures the voltage of the second coil instead of the current, then this voltage depends on the number of windings that each coil has. If both coils have an equal number of windings as shown in the figure, and if the coils are very close, then the voltage measured at the second coil equals that on the first during the time of changing magnetic field. If, however, the number of windings of the second coil is doubled, then the voltage is also doubled, if it is tripled, the voltage is tripled. The reason for this effect is similar to the situation with the seesaw. The longer end of the seesaw can exert a stronger force than the shorter. The electric field that is induced in the windings of the second coil represents the force and is the same throughout the second coil. If the second coil is longer (more windings) then the electric field adds up to a higher voltage at the ends of the coil. Of course, we cannot change the available energy, but we can change the available forces exactly as we could do mechanically by use of a seesaw or a hoist. Thus we can "transform" voltages, change them from larger to smaller, and vice versa by use of coils. This is done in cars to obtain the high voltages necessary for the spark plugs. The function of "transformers," that one often finds close to one's house, is also based on the principle of induction. They transform the currents and voltages from the high voltage of the power lines to the standard voltage in the household (110 V in the USA). The high voltages of the power lines must be lowered for household applications for safety and other reasons. The lowering can be achieved by using two coils with very different windings. Lets say that the coil that is connected to the high voltage state power line has 10,000 windings and the second coil that is connected to the household application has only 100 windings. Then the voltage for the household is lowered exactly by the ratio of the number of windings of each coil that is exactly by a factor of 100 in this example. The ignition coils in cars work the other way around. To create a spark one likes to increase the battery voltage by large factors.

Electric Motors

Another interesting application of coils and magnetism is the electric motor. Electric motors convert electrical into mechanical energy, for example, the rotation of some disc that drives washing machines, refrigerator pumps, cars, or even trains. There are two types of electric motors. One type uses alternating electric currents, that is, currents that change their direction. The power outlets in our households provide alternating currents (ac) that change their directions 50–60 times per second. Batteries provide only direct currents (dc) that flow in one direction only. One can, however, change a dc current into an ac current with relative ease. This is being done to create electric motors that can run on batteries. The principle of such motors is shown in Fig. 2.23.

A coil is mounted between two fixed magnets in such a way that it can rotate. One wire of the coil is electrically connected to one metallic half of a disc and the other wire to the other half of that disc. The two halves are separated by an insulating region, a region that does not conduct electricity. Two metallic springs slide around

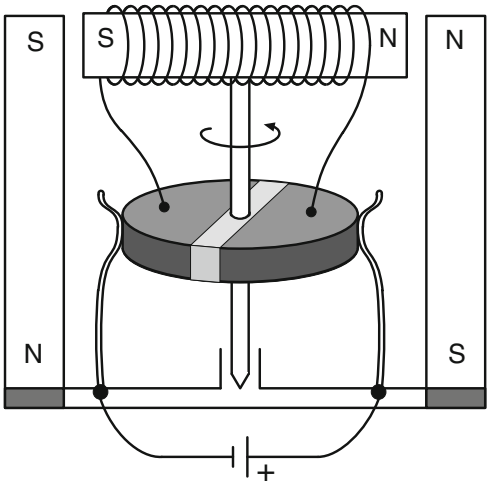


Fig. 2.23 A coil magnet is mounted between two other magnets (e.g., magnetized iron bars or additional coil magnets). The contact wires of the rotating coil are connected to the separated metallic surfaces of a disc that rotates with the magnet coil. The separation of the two metallic disc surfaces is achieved by a stripe of insulating material (*light grey*) between them. Metallic springs are sliding along the metallic disc boundary. As the disc turns, the current in the coil changes its direction and so does the magnetic field of the coil

the boundary of the disc and are connected to the battery. Assume now, that we start out with the position shown in Fig. 2.23. The coils north pole is close to the north pole of the fixed magnet (and the same is true for the south pole on the other side). Before the motion starts, the coil is turned just slightly as indicated by the arrow in the figure. The pairs of north and south poles repel each other, and the mobile coil starts moving away from the fixed poles, turning now more rapidly in the indicated direction. After half a turn, the north pole of the coil comes close to the south pole of the (left) fixed magnet and the south pole of the coil close to the (right) fixed north pole. Now the pairs of poles attract each other and the rotation continues. The rotation would then stop because of the attractions of north and south poles. However, as soon as a full rotation is complete, the springs that slide on the disc change to the other half of the disc. Then the direction of the current changes which in turn changes the magnetic field of the coil; the poles repel each other again and the motion continues.

From the description of this principle of electric motors, we can see that we need alternating currents through a coil to create rotating motion. If you use the ac power from the power outlet, the alternating current is available to start with, and the electric motor does not need any disc with sliding contacts. If you use dc electricity from a battery, then you need to use an electric motor that creates the alternating current from its own motion as described above. The springs that slide around the disc are the key to create that alternating current mechanically. There are nowadays also more modern ways, not involving mechanical parts but only electrical devices, to create alternating currents if needed.

2.2.3 Maxwell's Laws of Electromagnetism

Many of the initial experiments that are now common knowledge in the area of electromagnetic phenomena have been performed and explained by Faraday. Faraday was a self-made man with great genius and intuition. He and several other important contributors found the rules of electromagnetism and experimental ways to test these rules. Maxwell transformed these rules into mathematical equations called partial differential equations. As an example, we will discuss Maxwell's wave equation, the equation that describes all electromagnetic waves, in Sect. 5.3. Here we list the rules that are the basis for, and equivalent to, Maxwell's equations:

1. Electrical charges Q are the source of electric fields F that surround these charges. The electric field at a point in space is defined as the force per unit charge.
2. There exist no magnetic charges in analogy to the electric charges. The sources of magnetic fields are electric currents as illustrated in Fig. 2.21.
3. A magnetic field H encircles any electrical current.
4. An electric field F encircles any magnetic field $H(t)$ that changes with time t.

These are the laws that govern electric and magnetic fields and currents. Maxwell brought these laws into a strict mathematical form that was later improved by Hendrik Lorentz and Albert Einstein. Maxwell definitely had a type of fluid in mind, when he derived his equations, and we have outlined above how similar fluids and electricity behave. As a consequence, Maxwell also thought that the electrical phenomena take place in a fluid-like "ether" that fills all of space. Einstein's final version of electromagnetic phenomena is completely abstract and does not refer to any substance such as the ether. Einstein only talks about electric and magnetic fields, as well as their connections to each other and to electric charges and their motion.

2.2.4 Electromagnetic Waves, Wireless Communication

Wireless Communications

One of the most important consequences of Maxwell's rules is the possibility of wireless communications by use of electromagnetic fields (waves). Wireless communications are of great importance for technology and our daily life.

The first step, to understand how this particular application of electromagnetic waves works, is to understand how the waves can be created by use of the circuit shown in Fig. 2.24, a so-called resonant circuit. The circuit contains a capacitor which consists in essence of two parallel metal plates, an inductor, consisting of a wire coil, and a switch which is just a little piece of metal that can open and close the circuit. Assume that the capacitor is charged, which can be done by connecting it briefly to a battery. One is then left with positive charges on one plate and

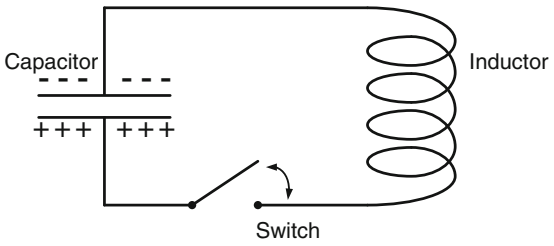


Fig. 2.24 Electrical resonant circuit with capacitor (parallel metal plates), inductor (coil), and switch (a wire connection that can be opened and closed)

negative charges on the other as indicated in Fig. 2.24. Up to this point, the switch is open. Now we close the switch. As soon as that happens, the electrons that have accumulated on the negative side of the capacitor start to move via the wires toward the positive end of the capacitor. This, however, means that an electrical current must flow through the coil, which in turn means that a magnetic field H is generated in and around the coil until the capacitor has lost all charge. Then the current stops, and the magnetic field starts to decrease, meaning that we have a time-dependent magnetic field $H(t)$ in and around the coil. According to Maxwell's rule (4) from above, the changing magnetic field causes an electric field and a current flow in the wire. This current flow occurs in the opposite direction (because the magnetic field is now collapsing) and has the consequence that the capacitor is now also being charged in the opposite direction. Then the capacitor starts to discharge, and the process continues until we arrive at the original situation, with one plate negatively and the other positively charged. If no energy is lost anywhere, this electromagnetic oscillation repeats itself over and over.

Heinrich Hertz understood from his detailed knowledge of the work of Maxwell, that an oscillating electrical circuit would lead to electromagnetic fields and waves in its neighborhood. To demonstrate the existence of electromagnetic waves in free space, Hertz constructed two identical resonant circuits of the type shown in Fig. 2.24 and placed them in two different rooms of his laboratory. He then showed that if he charged the capacitor in one room and started the oscillations, the oscillator in the next room would start oscillating by itself, even if the capacitor of this second circuit was not charged up by anyone. It was therefore clear that the electromagnetic oscillations were transmitted from one room to the next, just like oscillations of water waves or sound waves can be generated, transmitted, and detected. Wireless communication was born!

The formation and propagation of the electromagnetic waves created by Hertz is a consequence of Maxwell's rules (3) and (4). A graphical representation of the idea is shown in Fig. 2.25 that illustrates graphically how these rules lead to electromagnetic waves. Electric fields cause electric currents in Maxwell's ether (so-called displacement currents), and these currents are encircled by time-dependent magnetic fields. Time-dependent magnetic fields are, in turn, encircled by electric fields. In that way, an electromagnetic wave is created that propagates through space, as we understand it from the analogy to water waves and other types of waves. The velocity of electromagnetic waves, however, is extremely high, much higher than that of sound or water waves, and it took some time until it could be

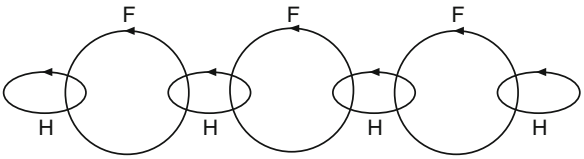


Fig. 2.25 Electrical (F) and magnetic (H) field lines encircling each other as demanded by the Maxwell–Faraday rules. Electromagnetic waves propagate through space in this way

reliably measured. The value of the velocity of electromagnetic waves in vacuum is generally denoted by c, and the measured value is $c = 3 \cdot 10^8$ m/s. This is an enormous speed. Such a wave travels from the earth to the moon in about one and a quarter seconds. All electromagnetic waves in vacuum travel with the velocity of light. The enormous speed is also an important factor for the use of wireless communications. It enables us to communicate almost instantly over large distances.

What is the frequency of the oscillations of a resonant circuit? The number of oscillations per second can be calculated from the following formula that we give without proof:

$$\nu = \frac{1}{2\pi\sqrt{LC}}, \tag{2.58}$$

C is the capacitance in Farads and L is the inductance in units of Henry (named after the American scientist Joseph Henry). The frequency is usually denoted by the Greek letter ν, and we have therefore also used that letter. The units of the frequency are s^{-1} (seconds to the minus one) because the frequency counts the oscillations per second. We have not discussed the unit Henry because this unit is not as easy to describe as, for example, the unit Farad for capacitance that was defined in Eq. (2.56). The interested reader is referred to the Internet with its detailed explanations. For very large capacitances and inductances one obtains (from Eq. (2.58) a few oscillations per second. One oscillation per second is also called one Hertz. For very small values of inductance and capacitance, the oscillation can be in the Giga Hertz range, which means we have 10^9 oscillations per second. Kilohertz (kHz) stands for 1,000 Hz, megahertz (MHz) for one million hertz, and one terahertz (THz) is 1,000 GHz. Modern wireless communications work mostly in the GHz range.

Why did it take more than hundred years after the discovery of Hertz to develop radio communications, and why do we only now have cell phones that can be used wirelessly almost anywhere? There are multiple answers to these questions. For one, the resonant circuits will not oscillate for very long because energy is lost to the propagating electromagnetic waves and also because of heat generation by the electrical currents in the resonant circuit (see Sect. 5.5.2). To maintain the oscillation, one needs some mechanism that feeds energy to the resonant circuit. This mechanism of feeding energy is called amplification, and the devices that feed the energy are called amplifiers. With help of such amplifiers, the oscillation can

go on forever. The Internet features many descriptions of amplifier circuits. The basic device that makes such amplifiers tick is the transistor of which we will hear more in Sect. 3.2. A second big hurdle to overcome was the need for very high frequencies of oscillation. One needs high frequencies for the following reason. If you think of digital information, information transmitted by zeros and ones, then, in order to create a one (a measurable signal), you need at least half a wavelength. You can think about this by observing water waves. The one could be the wave maximum and the zero the minimum. Thus, if you wish to transmit a gigabit of information, you need at least a frequency of one gigahertz. Electromagnetic waves with gigahertz frequency are called microwaves. Our cell phones work indeed in the gigahertz range, and it took the development of very special transistors that could amplify with such high speed. Other engineering advances were also necessary to make wireless communications possible. Antennas had to be developed that transmit (send) the waves from the basic oscillator circuit to the surrounding areas and the world, and other antennas that receive the transmitted signals in an optimal way and excite the receiving oscillator circuit.

Modern communication tools such as the Internet do use wireless communications via satellites and in Wi-Fi home networks, all working in the gigahertz frequencies. Some internet communications require much larger information transfer, measuring in the terabit (tera = 1,000 giga) range. To accomplish this, one needs communication by light waves. Light has a very high frequency and is, for this reason, capable of communicating a lot of information. It has the disadvantage that it is impeded and absorbed by obstacles that are in its way. This problem has been solved by using optical fibers. Optical fibers are made out of very transparent glasslike material, and they look like thin wires. There exist optical fiber connections between continents; the fiber cables are placed, for example, under water on the ocean floor between the USA and Europe. Of course, the fiber connection cannot be regarded as wireless anymore. The wires are just replaced by glass-type fibers. The communication is still achieved by the use of electromagnetic waves.

Range of Electromagnetic Waves: Radio Waves to γ-Rays

All electromagnetic waves are of similar nature. However, depending on their frequency, they appear very different to us. The generation of the electromagnetic waves is also very different for the different frequency ranges, and the resonant circuit described above is only useful for the "lower" frequencies, that is, for frequencies up to about 100 GHz. A given range of frequencies, such as the frequencies of the light of a rainbow, is called the "spectrum" of the electromagnetic waves. We will see in Sect. 2.5 that the frequency of the electromagnetic waves is closely related to their energy, and increasing frequency means increasing energy.

We have described already the waves of relatively low energy that are useful for radio and TV communications as well as cell phones. The resonant circuit of Fig. 2.24 can oscillate with a frequency as low as you like, if only capacitance and inductance are chosen large enough. For a 1 F capacitance and a 1 Henry inductance

the circuit oscillates precisely once every second and the frequency is then 1 H: $\nu = 1$ Hz. The electrical power that you can draw from a power outlet in your home oscillates at 60 Hz in the USA and at 50 Hz in Europe. Radio waves oscillate still much faster, typically in the range up to a few hundred megahertz (MHz). 100 MHz correspond to 10^8 oscillations per second. The cell phone interacts with its receiver station by sending out waves in the gigahertz (GHz) range. 1 GHz is 10^9 Hz. In this range we can use electromagnetic waves to generate heat in microwave ovens. Waves in this frequency range are called microwaves.

Electromagnetic waves at still higher frequencies, such as terahertz (THz) and above, are generated, for example, by heat and are emitted by a hot stove. One calls this type of radiation also infrared radiation, because its frequency is below that of visible red light that starts with frequencies of about $4 \cdot 10^{14}$ Hz. The light that our eyes can see ranges from red to violet which has a frequency of about $8 \cdot 10^{14}$ Hz. Beyond that frequency we have the so-called ultraviolet radiation and—still higher up in frequency—the spectral range of X-rays. X-rays are important for medical diagnostics and are used to perform CAT scans as described in Sect. 4.3. As one goes further up in frequency, one approaches the so-called γ-rays starting about at 10^{19} Hz. These latter rays of the highest frequency arise from events that involve the nucleus of atoms and radioactive processes. Often these processes are connected with the creation and also destruction of stars, and the γ-rays from stars are also called cosmic rays or cosmic radiation. They also are generated on earth, during the decay of radioactive atoms such as uranium or plutonium.

2.3 About Heat, Temperature, and Atoms

By now, we have learned about the laws of mechanics including the motion of planets, the laws of electricity including the workings of batteries, and about waves of all kinds including the electromagnetic spectrum. These are mostly phenomena that we can observe with our eyes and generally with our senses. We needed, however, some knowledge about the invisible atoms, electrons, and protons to explain, for example, the lithium battery. The existence of atoms is today generally accepted and agreed upon. Modern microscopes can produce computer images of atoms. Not too long ago, however, two famous scientists, Mach and Boltzmann, were fighting about the existence of atoms and Mach, and most people, did not believe in atoms. No powerful microscope was then available, and all information about atoms was very indirect. Boltzmann derived his ideas about atoms from phenomena connected to heat and how these phenomena could be explained by invoking atoms. Heat-related machines are very important in our daily life, ranging from car motors to jet engines and from refrigerators to air conditioning systems. This is what this section is about: how can we explain heat-related phenomena from the basic principles of mechanics and electricity, how can we use heat-related phenomena to engineer powerful machines, and how can we at the same time, find out more about atoms.

2.3.1 Heat and Temperature

It took scientists a long time to figure out what heat really is, and we are not going to discuss how they actually did find out. We present here only the final result: heat is nothing else but the kinetic energy of the atoms and molecules of the substance that exhibits heat. Consider a gas such as air as the substance. If we feel that a given volume of air is hot and another volume of air is cold, this means that the atoms and molecules of the hot air move with higher velocity than those of the cold air.

To be more exact, we need to have a method and an instrument to measure "hot" or "cold" or any "temperature." This method must provide us with a precise number that is the measure for hot or cold. Historically the temperature has been linked to the freezing and boiling of water. There are three temperature scales in common use. The temperature scale formulated by Celsius, and named after him, fixes its 0 point at the freezing temperature of water, while the water's boiling point is defined to be 100°C (in words: one hundred degrees Celsius). It is implied in this definition that the water is subject to a normal air pressure of precisely one atmosphere (see below). The Celsius scale is used predominantly in central Europe. The temperature scale that is used worldwide in science is derived from the fact that there exists an absolute zero of temperature, meaning that it is not possible to reach a lower temperature. The fact that the absolute zero cannot be reached by any procedure is known as the third law of thermodynamics (the first and second law are discussed below). This absolute zero is denoted by 0 K, where K is the abbreviation for Kelvin or degrees Kelvin. The absolute zero of the Kelvin scale equals −273.16°C which is far below the freezing point of water and even far below the temperatures of the north or south poles in coldest winter.

If you are living in the USA or in Great Britain, you may complain that the Fahrenheit scale was not mentioned yet, although this is the only well-known scale in your country. This is the third temperature scale of importance. You can easily find the conversion of all scales (Celsius, Kelvin, or Fahrenheit), as well as more about temperature scales and methods of measuring temperature, on the Internet. Here we need to be brief because we need to cover more modern topics in the limited space of this book. We also like to be scientific and therefore just use the Kelvin scale in the following. Except for the zero point, the Kelvin and Celsius scale are the same. You can obtain the Kelvin temperature from the Celsius temperature by just adding 273.16°.

2.3.2 Heat, Energy, and Avogadro's Number, Ideal Gases

The connection of heat and energy is illustrated in Fig. 2.26 that shows a movable piston within a cylinder that also contains a gas.

The principle is simple. If we increase the temperature of the gas in some way, it will expand and push the piston upwards. This push upwards can be used to perform

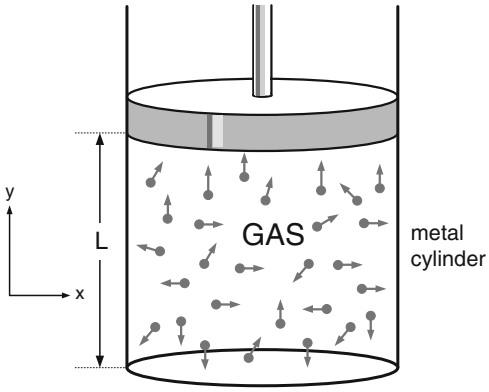


Fig. 2.26 A metal cylinder containing a gas and topped by a piston represents an important element of machines that convert some portion of heat into mechanical power

some work, for example, lifting a weight just as we discussed in Sect. 2.1.1. Because the performed work corresponds to energy and because energy is conserved, it follows that the energy comes from the heat and that heat is related to energy. We have already stated that heat corresponds to the kinetic energy of the atoms of the gas. This is also illustrated in Fig. 2.26 where we have included dots with arrows symbolizing atoms having a velocity in the direction of the arrow. We now derive from this basic knowledge the force that is exerted on the piston by the impact of the atoms. To do this, we need to remember Eq. (2.38) which says that the force F equals mass M times acceleration a; that is, $F = aM$. The acceleration that an atom encounters when hitting the wall is calculated in the following way.

Assume that the atom moves with a velocity v_y in the y-direction of the coordinate system which is the direction perpendicular to the piston as indicated in the figure. As soon as the atom hits the piston, it will be reflected and moves then in the opposite direction having the velocity $-v_y$. Thus the change in velocity is $2v_y$. To calculate the change in velocity per time period Δt, we need to know how often the atom hits the piston in that time period. This time period is given by

$$\Delta t = \frac{2L}{v_y}, \tag{2.59}$$

because any given particle must move up and down the full length L (from the bottom of the cylinder to the piston), in order to hit the piston just once for sure. From this relation we can calculate the acceleration a, which equals to the velocity change per time period:

$$a = \frac{2v_y}{\Delta t} = \frac{v_y^2}{L}. \tag{2.60}$$

The force exerted by the atom on the piston can then be calculated from Newton's law for the acceleration:

$$F = Ma = M\frac{v_y^2}{L}. \tag{2.61}$$

This is the time-averaged force exerted by a single atom roaming around in the cylinder and hitting the piston. The number of atoms in a volume of a few thousand cubic centimeters is incredibly large. Therefore, huge numbers of atoms hit the piston at any instant of time. Their velocity and energy depends on the temperature T (that we measure in Kelvin). The force exerted on the piston is, therefore, a statistical average and sum of a large number of atomic forces. This average force can be calculated if we know the relationship of Mv_y^2 to the temperature. It is intuitively obvious that the average kinetic energy in the y-direction, which equals $\frac{Mv_y^2}{2}$, must be proportional to the temperature. If we denote the constant of proportionality by k_B, we have

$$M < v_y^2 >= k_\mathrm{B} T. \tag{2.62}$$

Here the symbol $< \ldots >$ indicates that a statistical average over many events has been taken. The atoms move actually all with different velocities, but if we average over all these velocities, then this is the result we get for the average force that one atom exerts on the piston:

$$< F >= \frac{k_\mathrm{B} T}{L}, \tag{2.63}$$

while N atoms exert then an average force of

$$< F >= N \frac{k_\mathrm{B} T}{L}. \tag{2.64}$$

The constant k_B is now called Boltzmann's constant, a number that you can look up, together with its history, on the Internet. It was actually named by Max Planck, who wished to honor Boltzmann's pioneering work. Boltzmann himself was too absorbed with his fight to validate atomic models for gases such as the one we have just described. He never pursued any measurement of the average kinetic energy of atoms or molecules himself.

The pressure p exerted on the piston is defined as average force divided by the area A of the piston. Thus we have

$$p = \frac{< F >}{A} = N \frac{k_\mathrm{B} T}{LA} = N \frac{k_\mathrm{B} T}{V_\mathrm{ol}}, \tag{2.65}$$

where $V_\mathrm{ol} = LA$ is the volume of the gas in the cylinder. We can reformulate the last equation to obtain:

$$pV_\mathrm{ol} = N k_\mathrm{B} T. \tag{2.66}$$

This is the basic equation for an ideal gas. Ideal means here that we have assumed that the atoms interact with each other exactly the same way billiard balls do in their elastic collisions. We have described these collisions by the equations for energy and momentum conservation in Sect. 2.1.2, and we have used the essence of these equations in the derivation of Eq. (2.66). If the atoms did interact in some other

way, the resulting equation for pressure and volume of the gas would be different. For example, the atoms could attract each other over larger distances by van der Waal forces (see below). That would make things more complicated. Indeed, as we will discuss below, real gases do have such atomic interactions, and the above equation for ideal gases is only approximately valid. However, the equation agrees incredibly well with the experiments as long as the temperature is not too low. This agreement represents a great success of the methods of science. Just remember that the laws that we used were derived from the way stones are falling and planets are moving. Now we have used these laws to calculate the pressure on a piston; a pressure that is caused by an enormous number of atoms swarming around in an otherwise empty space. This statistical bombardment of the piston by atoms is what causes the pressure. Equation (2.66) is one of the first equations of physics that involves statistics and probability, because it involves averages over large numbers of particles. It tells us that the pressure that we measure is just some average value caused by atomic bombardment. On a fine scale there must therefore be fluctuations, because of the fluctuations of the bombardment of the piston with different numbers of atoms.

Pressure is usually measured in "atmospheres." One atmosphere is the force per unit area (per square centimeter) that is exerted by the air above as measured at sea level. So, consider that you have a container of gas with a pressure of 1 atmosphere closed off by a piston on top and you are standing at the sea side. The piston then will not be moving, because the air from above bombards the piston on the other side and the pressure from above and below balance each other exactly. If you evacuate the container, then the piston will be pressed down by the significant force of the air above it. The pressure of the air at sea level corresponds almost exactly to the pressure that 10 m of water above the piston would exert on the piston. Therefore, if you dive 10 m below sea level, the pressure is about 2 atmospheres, one arising from the air, the other from the water above you. For a more detailed definition and other units of pressure, consult the Internet.

The number N of Eq. (2.66) of atoms in a gas can be determined by using our knowledge of Galvanic elements. Consider Eq. (2.49). This equation tells us how many lithium atoms are freed at one side of the Galvanic element for each electron that we supply in the form of electrical current when we recharge the battery. Therefore, we can count the number of charges that we supply through the electrical current during a given time period, and that number equals the number of atoms that we deposit at that side. Lithium is actually difficult to obtain in gas form. However, we can do the same experiment by using hydrogen. Then we can generate, for example, about one gram of hydrogen and then count the number of electrons that generate the hydrogen by measuring the electric current. This number is equal to the number N_A of hydrogen atoms in about one gram of hydrogen and is called Avogadro's number, which is

$$N_A = 6.0221415 \cdot 10^{23}. \tag{2.67}$$

You may have noted a little inconsistency here. We said "about one gram of hydrogen" and then gave Avogadro's number to seven decimals. The reason behind

this is that nature does not supply us with hydrogen (or any other element for that matter) in its purest form, with every atom being the same and consisting of precisely one proton and one electron. There are usually also heavier or lighter atoms present in any given material with the same number of electrons and protons per atom. The difference in weight arises from an additional particle that is called a neutron. This particle does not change the electrical properties of the atom because it is neutral. It does change, however, the weight. In the case of hydrogen the weight almost doubles when a neutron is added, and one calls the atom then deuterium. Deuterium is in our drinking water and everywhere, but only in small quantities (about 1 Deuterium atom for every 6,000 hydrogen atoms). Deuterium is called an isotope of hydrogen. The topic of isotopes is an interesting one for advanced projects and Internet searches. Avogadro's number was actually determined before neutrons and isotopes were known. We proceed in the following by ignoring the weight corrections due to isotopes and just use the word "about" to indicate that the actual weights of gas atoms that are measured in nature may be slightly different because of isotope effects.

If we concatenate Avogadro's number with atoms other than hydrogen, we obtain different weights measured in grams. For example, we obtain about 4 g of helium, 6 g of lithium, 12 g of carbon, and so forth. The reason why the weight is close to a whole number is the following. The atoms consist of electrons that weigh very little and of neutrons and protons that have almost equal weight. Hydrogen has one proton and Avogadro's number of hydrogen atoms weigh about one gram. Therefore we can conclude that the weight for the other atoms indicates the sum of protons and neutrons. The number of protons and neutrons is equal for the above-mentioned atoms. We can deduct from this that helium contains two protons and two neutrons, lithium contains three, and carbon six protons and six neutrons. The atoms also contain as many electrons as they contain protons because the total charge must be zero as we know from our discussions of Galvanic elements. Thus, once we have a hypothesis that atoms exist and that they consist of protons neutrons and electrons, we can deduce the number of all these constituents from rather simple measurements. Of course, the inverse process, the actual deduction of the number of atoms, electrons, protons, and neutrons from measurements, has been much more complicated and has required great ingenuity from the scientists that went that arduous path. To describe how they actually found all of these facts would lead us too far away from our major objectives.

Avogadro's number is an incredibly large number. 10^{23} is a number with 23 zeros. I would like to illustrate this by the following stunning example. Think of the great roman leader Julius Caesar after whom the month of July is named. In a few days of breathing, Caesar was inhaling oxygen and exhaling, by easy estimates, more than a few grams of carbon dioxide molecules and therefore at least Avogadro's number of them. Lets assume now that these atoms of Caesar's breath have distributed themselves in the last 2,000 years in equal amounts all over the world's atmosphere. Now we calculate the volume of the atmosphere of the world. The radius R of the world is $6.37 \cdot 10^6$ m. The surface area of a sphere with that radius is $4R^2\pi \approx 5.1 \cdot 10^{14}$ square meters (m^2). Because the atmosphere is approximately

10,000 m high, that gives us an air volume V_{air} of about $V_{\text{air}} = \approx 5.1 \cdot 10^{18}$ cubic meters (m^3). In one cubic meter of air we have therefore about $\frac{6.02 \cdot 10^{23}}{5.1 \cdot 10^{18}} = 1.2 \cdot 10^5$ carbon dioxide molecules that Caesar had in his lungs. Because we breathe a cubic meter of air within an hour or so, we breath every hour more than 100,000 atoms that also Caesar was breathing. The story gets a little funnier, if you realize that the gas exhaustion of Caesar did also include gases that Caesar did not just exhale from his lungs, and, in addition, we would also breath gases from his dog, if he had a dog.

There is a great lesson to be learned here about dealing with atoms or molecules. Because there are so many atoms in every reasonably large volume of our surroundings, our body will encounter numbers of atoms that appear to be large, as the 100,000 atoms of Caesar, but are, in fact, very small compared to Avogadro's number. It follows therefore that our body can usually deal with such numbers without problems. For example, if we inhale air through a piece of garden hose made out of rubber, then we will breath in thousands and probably millions of sulfur atoms because rubber contains sulfur. This will not influence us, however, in any major way. If it did, a gardener would live shorter than other people because of just handling the garden hoses. Current science permits us to measure incredibly small quantities of atoms. In some cases, measurements are sensitive to a few atoms. Only in rare cases will 100,000 or fewer atoms hurt our body. There are, of course, poisons that are so potent that they may do harm even in small quantities. One needs therefore to apply STEM reasoning if one deals with TV—or newspaper—reports that traces of a certain substance such as sulfur have been found above normal levels in living tissue. Maybe the tissue has just touched a garden hose. Then, there is no need to get excited about small traces of ordinary chemicals. Of course, with potent poisons, one cannot be careful enough.

We can see that the pressure that a gas excerts on a simple piston, together with the idea of atoms, permits us to discover a lot about the atoms. It so happens that pistons driven by a gas are the building blocks of very important mechanical motors including the steam engine and the engines of cars. Both types of motors are based on the fact that heat is a form of energy.

2.3.3 Steam and Combustion Engines

An engine is defined as a machine or instrument that turns energy into mechanical motion. We just discuss the principles of a few such machines that are extremely important for our daily life and power our cars and airplanes.

Steam Engines

The steam engine is of historical importance and is still in use for some applications. For the steam engine, the gas of operation is, of course, steam. Steam is nothing else but water heated so much that it changes from its liquid form to become a gas.

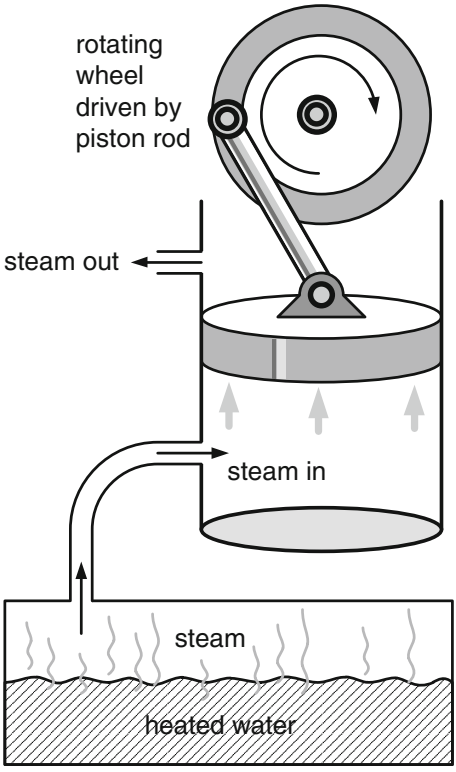


Fig. 2.27 Principle of the steam engine. A container with water is heated so that pressurized steam is generated. The steam is transferred to a cylinder with piston. The pressure pushes the piston upward and the piston is connected to a disc (wheel) that turns. When the piston passes the opening at the left of the cylinder, steam is released while the disc keeps turning because of its inertia and pushes the piston below the release opening. Then the steam pressure increases again and pushes the piston upwards. The process continues and the disc keeps turning

The water molecules are moving so fast that they cannot stick together as they do in the liquid but move farther apart from each other. Water molecules are composed of hydrogen for which we use the chemical symbol H and of oxygen for which we use the chemical symbol O. Each water molecule consist of two hydrogen atoms and one oxygen atom. It is therefore represented by the symbol H_2O. In the liquid, these molecules are attracted to each other and stick very closely together for reasons that we will discuss below. The principle of the steam engine is shown in Fig. 2.27.

In a cylinder, a piston is pressured upward by the hot steam coming from a boiler that is heated so that water evaporates. The piston is connected by a metal rod to a disc (wheel) that can rotate around an axis. The connections of the metal rod to both the disc and the piston are flexible. As the piston pushes upward the disc starts to turn and continues turning until the steam is released through an opening in the upper portion of the cylinder, and the rotating disc moves the piston back to the original position. Then the process repeats itself and the disc keeps turning. The power of the turning disc can be used to drive a train or other useful machinery. Steam engines have revolutionized mankind. Since their invention, the power of the muscle could be replaced by the greater power of machines. Factories using steam engines where built all over the world, and railroads have transported people and loads around the country much faster than horses could. Steam engines are

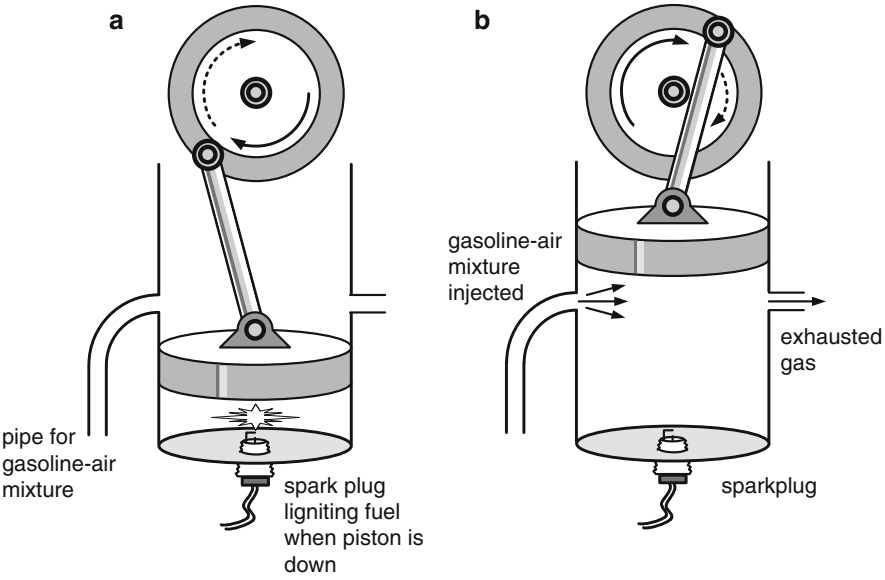


Fig. 2.28 Principle of the two-stroke internal combustion engine, an engine that still drives lawn mowers and other vehicles. (**a**) In stroke one, a spark plug ignites the mixture of gasoline and air that has been injected into the metal cylinder and compressed. The electrical spark is generated at the right moment when the piston is close to its lowest point. The ignited hot gas pressures the piston upwards. (**b**) In stroke two, the hot gas is exhausted through the upper opening on the *right* side of the cylinder, while a new air–gasoline mixture is injected through an opening on the *left*. As in the case of the steam engine, the piston is connected to a rotating disc that continues to rotate through its inertia and pushes the piston with the gasoline–air mixture down. Close to the lowest point of the piston, the spark plug is ignited again and the process is repeated keeping the disc spinning

used in modern nuclear power plants and supply a significant percentage of the worlds energy. The importance of steam engines for mankind would certainly justify a detailed description of the more sophisticated modern implementations, and the reader is encouraged to surf the Internet to find these descriptions.

Combustion Engines

The next revolution of this type was caused by the invention of the so-called "internal combustion engine." The principle of this engine is similar to that of the steam engine. However, instead of steam, a combustible gas, usually a mixture of gasoline (or alcohol) and oxygen from the air, is injected into the cylinder and then electrically ignited at the right moment by a spark plug. The exploding gas pushes then the piston upward. This is illustrated in Fig. 2.28.

This machine is called also a two-stroke engine, because there are essentially two positions that are important for the machine to run. First, when the piston

is compressing a mixture of gasoline (or alcohol) and air, a spark is generated electrically at the bottom of the cylinder by a spark plug. The gas mixture combusts then and the hot gas generates a high pressure. This fact can be seen from Eq. (2.66) that tells us that the pressure rises when the temperature rises while the volume stays at that moment the same. Thus, second, the piston is pushed up with great force until it gets to the opening in the cylinder where it releases the combusted gases. After this release, new air–gasoline mixture is sucked in (or injected), and the piston is pushed down again by the turning disc. Two-stroke engines are still in use, for example, for mopeds, mowing machines, and other garden machinery. They were even used for some cars. Most of the cars have more complicated four-stroke engines, but the principle is the same. There are various videos with moving pistons on the Internet (Wikipedia) that help with a more detailed understanding.

Rocket Engines, Jet Engines

Rocket engines are, at least in principle, the simplest of all engines. They consist in essence of a cylinder that is narrow at one end and contains a chemical that is usually solid or liquid and that can rapidly combust. The process of combustion (burning) transforms the solid or liquid into a hot gas that occupies a much larger volume. As discussed previously, a solid or liquid has a much higher density than a gas. The gas therefore leaves the cylinder under pressure with a very high velocity. To understand the consequences, we remember our discussion of billiard ball movements and collisions in Sect. 2.1.2. There we have learned that the total momentum of all billiard balls is conserved and must, therefore, stay the same in all processes that might happen. We treat now all the atoms of the rocket as billiard balls. To simplify the problem, we assume that the rocket is somewhere in outer space with no forces acting on it. Before the ignition of the fuel, the rocket stands still which means that the total momentum is zero. After the ignition gas streams out of the rocket with high speed in the backward direction as shown in Fig. 2.29. We chose the direction of the out-streaming gas as the negative direction of our coordinate system. The gas has, therefore, a negative momentum $-M_G v_G$, with M_G being the mass of the expelled gas and v_G being its velocity. The rocket has a different mass M_R, and its velocity at the beginning is zero. However, we can calculate its velocity v_R after ignition from the fact that the total momentum must always stay zero. Therefore, if after some time all the exhausted gas has a momentum of $-M_G v_G$, we must have

$$M_R v_R - M_G v_G = 0, \tag{2.68}$$

where $M_R v_R$ is the momentum of the rocket at that time. From this we immediately obtain

$$M_R v_R = M_G v_G \tag{2.69}$$

and

$$v_R = \frac{M_G v_G}{M_R}, \tag{2.70}$$

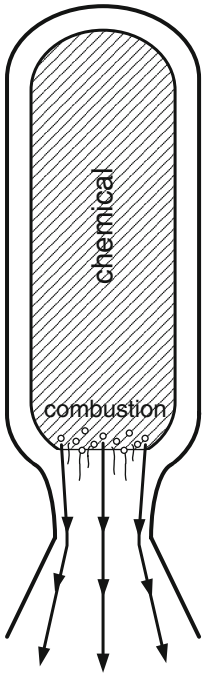


Fig. 2.29 Principle of a rocket engine: A sturdy steel cylinder is filled with a chemical that creates highly pressurized hot gas when burned. One end of the cylinder has an opening through which the gas is expelled

which means that the rocket moves with the positive velocity v_R, i.e., it moves in the positive x-direction opposite to the exhaust gas. Actual rocket science is based on this principle of momentum conservation. The art of rocket engineering is, of course, to get the rocket exactly where one wants it to go, as, for example, into an orbit around the earth. One therefore needs to include the gravitational forces and accelerations and thus needs to solve Newton's equations. Such a solution, including the gravitational forces of earth, other planets, moon(s), and sun can only be accomplished by use of computers.

The workings of jet engines are a little more involved than that of rocket engines. The jet engine principle is illustrated in Fig. 2.30. Jet engines are used to power airplanes and use fuel that is similar to the gasoline that is used in cars. In fact the fuel is usually a liquid called kerosine or a kerosine–gasoline mixture. The kerosine–gasoline mixture does need the oxygen of air in order to burn. The air is supplied in jet engines by a fan (the fan on the left side shown in Fig. 2.30). That fan sucks air in and compresses it into cylinders that resemble rocket engines, two of which are also shown in Fig. 2.30. These rocket-engine cylinders have an opening through which the pressurized air enters. There is no solid rocket fuel inside these cylinders. Instead, the gasoline–kerosine mixture is injected through a pipe as also shown. The air–gasoline–kerosine mixture is then ignited (e.g., by electric spark), and the combusted hot gas is propelled toward the right big opening of the cylinder and hits there a second fan and drives this fan to high rotational speed. The second fan is, in turn, connected by a metal rod (center axis) to the first (left) fan that sucks the air in.

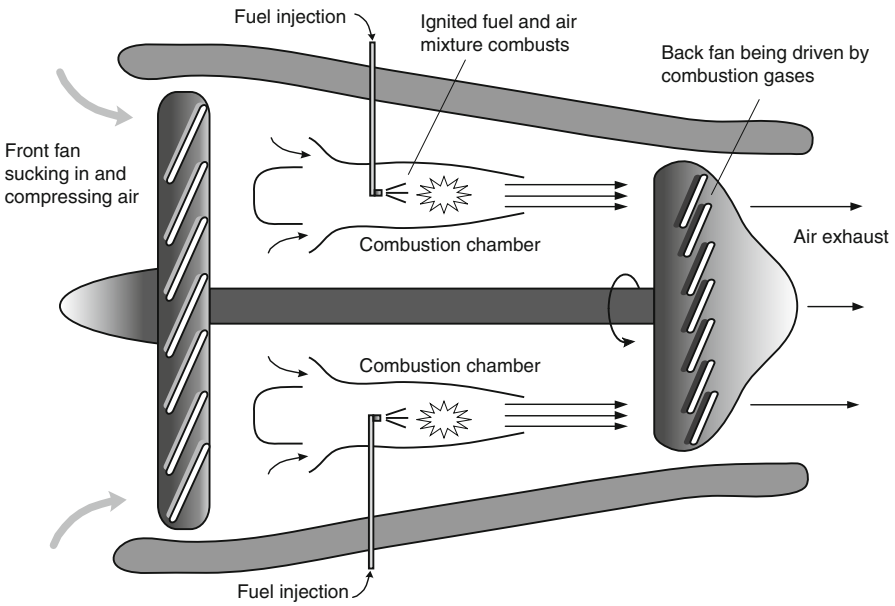


Fig. 2.30 Principle of a jet engine: A front fan is connected by a cylindrical rod to a second fan at the engines back. The front fan takes air into the combustion chamber and compresses it. Fuel is injected into these combustion chambers and ignited causing a jet of compressed air and combusted fuel to accelerate the back fan that in turn by the connection rod drives the front fan

All air is subsequently blown out at the back after being fully combusted. As is well known, such jet engines are very powerful and can lift huge airplanes. The engineering principles, that make such engines work, are the physics principles of the rocket engine (momentum conservation) plus another principle that we have not discussed yet. This second principle is the feedback between the front (left) and the back (right) fans. The back fan drives the front fan that sucks the air in and in this way pressures air into the rocket cylinders that pressure combusted air onto the back fan. The action of the two fans feedback on each other. Feedback as it is commonly termed is very important in many engineering applications. It is also important that the feedback can be controlled so that the desired power can be reached. This control is exerted by the amount of fuel that is injected. That amount is, in turn, controlled by the pilot of the plane who operates the levers that control the fuel and therefore the thrust of the engines.

2.3.4 Refrigerators and Nonideal Gases

Steam and combustion engines work at high temperatures, and this usually means that the gases that are involved behave like ideal gases. The reason is that, at high

temperatures, the atoms or molecules move very fast and spend only a short time close to each other. Their interactions are thus limited to collisions of short duration. This type of interactions randomizes the direction of the velocities. The molecules and atoms behave as if they would be fully elastic billiard balls.

At lower temperatures, the velocities of the atoms and molecules are much smaller, and atoms and molecules reside next to each other for longer time periods. As a consequence, the interactions between atoms and particularly between molecules, influence the behavior that is seen by the outside world. The case of steam is a good example. If one lowers the temperature, steam condenses into water and becomes a liquid instead of a gas. Liquids are essentially not compressible, and their volume stays almost constant. Equation (2.66) is then not valid. However, the equation that describes the gas becomes already nonideal before the gas liquifies, because of the stronger interactions of the molecules at the lower temperature. These molecular interactions are a very important component for many processes in nature. They explain why geckos can run on ceilings and why water moves up through the wood of trees, from the roots to the leaves. We describe these interactions therefore in some detail and first explain how refrigerators work on the basis of these molecular interactions. Subsequently we give a brief overview of how these interactions form the basis of many processes in biology.

The physicist Johannes van der Waals was the pioneer who discussed the molecular interactions in gases. We mention his name to facilitate Internet searches for those who wish to know more about this area. The forces that are the origin of these interactions can be explained well by the example of the water molecule H_2O. Both oxygen and hydrogen are, when viewed from some distance, electrically neutral. This is because hydrogen consists of one negatively charged electron and one positively charged proton. Oxygen possesses 8 negative electrons and 8 positively charged protons in its atomic core or nucleus. When hydrogen and oxygen approach each other closely a molecule of water forms as indicated in Fig. 2.31.

The formation of molecules is described in detail in Sects. 2.4.2 and 2.5.3. For now it is sufficient to know that oxygen "likes" the electrons more than hydrogen. The consequence of this fact is shown in Fig. 2.31: the side of the molecule that contains the oxygen nucleus is more negatively charged and the side with the hydrogen more positively charged. One says that such a molecule is a dipole, meaning it has a positive and a negative side. Such a dipole likes to attract other dipoles in the way shown in the figure with positive sides pairing with negative sides and vice versa. This type of attraction and the corresponding forces, are a typical example of the forces between molecules that van der Waals had in mind. He knew that it takes energy to break up such clusters of molecules that are connected by the dipole forces and to get them farther apart. We now discuss an example of the consequences of such a breakup of van der Waals forces.

Consider the metal cylinder shown in Fig. 2.32. The wall in the middle has a small hole through which high pressure water vapor (moist air) or steam is supplied, while to the right, an outward-moving piston lowers the pressure considerably (creates a vacuum). The pressurized water vapor expands as soon as it enters the right half of the cylinder. Therefore, the water molecules move away from each other. To do so,

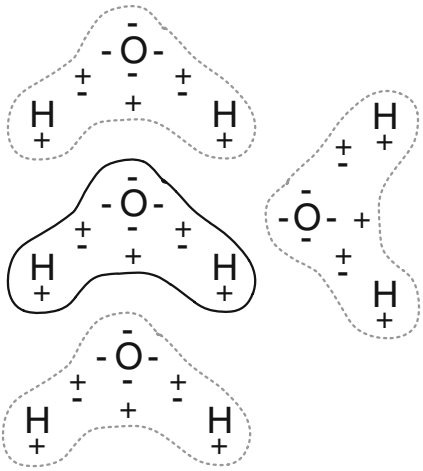


Fig. 2.31 Schematic drawing of the charge distribution of water molecules. The oxygen *O* "hogs" more of the negative charge of the H_2O molecule than the proton of the hydrogen *H* atom does. The water molecule has a shape similar to that shown in the figure. There are more negative charges "swarming" close to the oxygen, while the hydrogen side is more positively charged. If other water molecules are present, as they are shown (*dashed*) in the figure, then these molecules are attracted in such a way that the negative oxygen stays close to the positive hydrogen of the other molecules

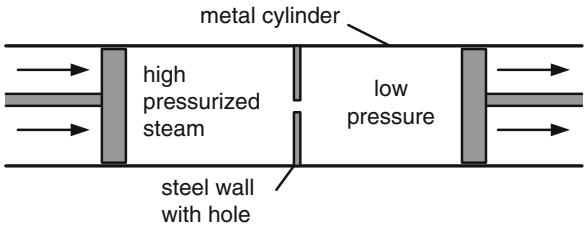


Fig. 2.32 Cylinder with piston (*left*) pressing water vapor through a small opening. A second piston moves to the *right* faster than the first one and thus lowers the pressure in the right half of the cylinder. The pressure lowering leads to a cooling of the right side. This type of cooling effect is used in refrigerators

the van der Waals forces that attract the molecules need to be overcome, and that does take energy. This energy can only be taken from the heat content of the water vapor (steam). Therefore, the expansion of the gas slows the water molecules down, which means the water vapor is cooled. One can achieve the same effect with gases other than water vapor or steam. One can use, for example, propane gas and start with pressurized propane at room temperature. Then, the expanded propane cools below room temperature. You can observe this effect if you have a gas grill supplied by a propane container. Open the valve on top of the propane container and ignite the propane as usual to operate the gas grill. The fire heats now the top grill and the whole equipment gets a little warmer. However, just at the top of the propane

container, where the gas streams out and expands, you can clearly feel a significant cooling effect. This is the principle of how a refrigerator works, the only difference being that the gases of the refrigerator are not ignited anywhere, and gases other than propane are used that have a higher efficiency for cooling.

In this way cooling can be achieved in a closed space (e.g., the inside of a refrigerator). The gas itself is warmed up somewhat by the process of cooling the refrigerator's inside and then transported to the outside. There the gas is compressed again which warms it up. Fans or some other cooling equipment bring it back to room temperature. Then again, the gas is decompressed (expanded) in the tubes inside the refrigerator and cools them down. The net effect is that heat is transported out of the refrigerator to the kitchen. In addition work needs to be performed (energy used) for the compression process that also heats the kitchen. This process is called the refrigeration cycle. The inverse process is also possible. One can cool down (as just described) some space outside a house and heat up the inside of the house. A machine that accomplishes that is called a heat pump. Heat pumps are more efficient in energy use than furnaces that just burn gas, wood coal, or oil (or anything else). The reason is that a heat pump moves the heat and uses only the energy necessary for that moving. The process is of importance for possible reductions of energy usage and a way of heating that is environmentally friendly.

Thus, the physics of gases is important for a number of applications including car engines and refrigerators. At the same time one can learn a lot about atoms by considering the detailed mechanisms of these machines. It is quite common in the history of science that engineering applications that derive from science have a positive feedback and enhance and extend the scientific work of a given area. Van der Waals' work represents a great example. He was mainly concerned with deviations from the equation for the ideal gas. However, the forces between the molecules are of great importance in many areas of science and engineering.

Interactions between positive and negative charges were already well known before van der Waals' work. He pointed toward a new phenomenon: even when molecules are overall neutral, i.e., have the same number of positive and negative charges, considerable interactions between them are still possible because of the so-called dipole forces that are illustrated in Fig. 2.31. These forces are strong over short distances and act only if the molecules are close to each other. Forces of this kind also occur between liquids and solids and between two solids. A very important effect of this kind is the so-called capillary action that is illustrated in Fig. 2.33.

A capillary is just a very thin tube. When such a tube is immersed into a liquid, two effects can happen. If the liquid and the tube material attract each other, because of the existence of electrical forces as we just have described them, the liquid "creeps" up the tube walls and is therefore pulled up higher. If the tube is very thin the liquid can move up many meters (yards) high. This is actually one of the physical mechanisms that helps plants transfer water from the ground to their leaves. Repulsive forces between liquid and capillary wall are also possible and also illustrated in the figure. Such forces are, for example, observed between mercury (which is a liquid metal at room temperature) and glass. The explanation of these repulsive forces is not as straightforward as the explanation

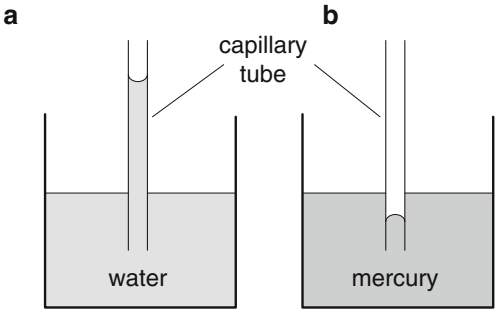


Fig. 2.33 Behavior of liquids in a capillary (thin tube). Part (**a**) shows the result if an attractive force exists between liquid and the inner wall of the capillary tube. Such an attraction exists between certain materials (e.g., glass) and water. The water is then pulled upwards. For very thin tubes, the water may be pulled up several meters (yards). Part (**b**) shows the result for a repulsive force as it occurs, for example, between mercury and glass: the mercury is pushed down. Note that the curvature of the liquid that is shown within the capillary occurs also at all other liquid-glass boundaries but is not shown

of the van der Waals forces was. The general theory that explains all of these forces, attractive and repulsive, was pioneered by Hendrik Casimir. Van der Waals and Casimir forces are of central importance in nano-science and nanoengineering, because the attraction and repulsion of nanostructures is of great importance for the workings of nanomechanical machines. Miniature laboratories made on "chips" may have movable parts that can get stuck because of attractive van der Waals forces. The interested reader is referred to Sect. 3.4 and to the Internet.

We finish this section with the following remark. Electrical forces such as van der Waals forces and Casimir forces are much stronger than gravitational forces such as the attraction of massive bodies by the earth. This is why water can be transferred against the gravitational pull to the top of a high tree.

2.3.5 The Random Motion of Atoms and Laws of Thermodynamics

Heat-related phenomena have a great significance for both science and engineering. From the engineering point of view, it is important to obtain the utmost mechanical energy from steam and combustion engines with the least amount of fuel. It is clear that we wish our cars to drive the largest distance with the greatest possible flexibility and power. The achieved miles or kilometers per gallon (or liter) of gasoline become of greater concern the more cars we use and the more gasoline is consumed. Gasoline is usually made from oil, and the oil has formed over millions of years from life-forms such as algae. Oil and gasoline have, therefore, received the name "fossil fuels." Fossil fuels are not easily replaceable, because their formation

takes so long. Mankind has learned to create oils, alcohol, and other fuel types more quickly by using plants, algae, and bacteria that grow now and have not just grown in the past. These fuels are called renewable fuels, because they can be generated over and over, as our needs require. As these lines are written, however, the renewable fuels are more expensive to generate than fossil fuels are, and fossil fuels are therefore predominantly used, in spite of the fact that they cannot be replaced. We return to this topic in Sect. 4.1.4, and just note here that the supply of sufficient fuel energy to the population of the world is a very complicated and important problem. Therefore we must conserve fuels as much as possible and use engines (motors) that have the greatest mechanical energy output for the least fuel input. If we look at the engines that we have discussed above, however, we see that these engines are far from ideal. The two-stroke engine, for example, exhausts even fuel that is not completely burned, and all of these engines exhaust some hot gas, gas that carries still a lot of energy with it. The question therefore arises whether more ideal engines can be built, and engineers are continually thinking about this question. For example, there are investigations in progress about whether we can do better with electric motors. This is a great area for future STEM experts.

There is a fundamental limit to the efficiency of all engines that use heat in some form. This limit was investigated by the French physicist Nicolas Carnot. The limit is in essence based on the fact that heat is the random motion of atoms. Carnot found that we are never able to utilize all the energy stored in heat, because of this randomness of atom motion. What we wish to produce with all engines is, at the end, some mechanical power, as that of a moving piston. The pistons of engines move in a certain direction, and the gas that drives the jet engines moves also in a given direction, while heat is based on the random movement of atoms in all directions. In all the heat-based engines, we always lose some energy to the random motion of atoms and molecules. This fact is expressed by one of the basic laws of physics that is called the second law of thermodynamics (the dynamics involving heat). The second law says it is not possible to construct an engine (such as the engine of a car) that has no other effect than creating mechanical motion, while deriving the energy by cooling down some substance.

It is thus impossible that a stone just cools down and flies up into the air by using its own heat energy. This is, of course, rather obvious to us. Note, however, that the second law of thermodynamics has been checked and proven experimentally over and over and is considered as valid as the first law. The first law of thermodynamics has been discussed already above in various ways. The first law states, in essence, that heat is just a form of energy and that the sum of the various forms of energy, electrical, mechanical, and heat energy, stays constant in any closed system. This first law tells us, as we have learned on other occasion, that energy is conserved. It is therefore impossible to run a machine without supplying the energy that is produced by the machine in a different form (often mechanical or electrical). Usually the energy is supplied to the machine through some form of fuel.

It is very important to realize that we need machines to sustain human life in modern ways. We need therefore energy in the form of fuel or electricity or heat. It is also important to realize that running these machines always involves creation

of some random form of energy. It is impossible for us to just cool down a heat reservoir and draw mechanical energy and do nothing else to the environment. The second law tells us that, if we wish to live and use energy, we must perturb the environment by creating heat and exhaust gases. All we can do is search for the most efficient machines that have minimal negative influences to our environment. This is a very complicated problem indeed, and one worthy for scientists and engineers to investigate probably for as long as humans populate the planet.

2.4 Chemistry and Quantum Mechanics

Chemistry and quantum mechanics are sometimes seen as separate subjects. However, chemistry had a unique influence on the discovery of quantum properties, properties that are connected to the fact that the world that surrounds us is made of elementary building blocks such as atoms, electrons, and protons. These building blocks are typically encountered and measured as a whole a so-called "quantum" and cannot easily be divided into smaller entities, although such a division is possible for the atoms and, as we will see, even for protons. Chemistry has discovered and researched basically all possible types of atoms and the way these atoms form bonds and molecules. All the materials that we encounter on earth can be understood that way, and chemistry is therefore a central science of great reach and importance. Quantum mechanics developed from the findings of chemical science and came up with a method to compute all the properties of atoms, and molecules, particularly the important energies that are involved in molecule formation. Very important steps were added by the Austrian physicist Erwin Schrödinger who found a wave equation that accomplished all of this quantitative understanding of molecules. Quantum mechanics developed then further and gave new meaning to the particles of chemistry. In particular, quantum mechanics tells us that all the particles that we can identify in nature have also properties that are typical for waves. There are some scientists who call all these entities "wavicles," and wavicles make up our world.

These introductory sentences, however, have brought us far ahead of the historic developments that have first provided an understanding of the atomic composition of our surroundings. We start the next section, therefore, with the discoveries of chemistry that have shown that the materials around us consist of different atoms or compositions of atoms (molecules), and then explain the properties of these atoms and of important molecules in terms of relatively simple rules. Only the last section deals with quanta in a more general way. There we discuss photons, the quanta of light, the emission and absorption of photons by atoms, and the energy range that is characteristic for the light emitted from atoms and molecules (the so-called spectrum). The principles of quantum mechanics are also used to derive the energies of electrons and the characteristic energies for the bonds between atoms in molecules.

1 H							2 He
3 Li	4 Be	5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar

Fig. 2.34 The first 18 elements (atoms) of the so-called periodic system of elements, starting from the lightest (hydrogen) and ending with the heaviest (argon). Atoms arranged in the same column (such as H, Li, Na, or C, Si) have similar chemical properties. For example, He, Ne, Ar are called noble gases, have low chemical reactivity, and are odorless and colorless. The names of the atoms corresponding to the symbols are (H) hydrogen, (He) helium, (Li) lithium, (Be) beryllium, (B) boron, (C) carbon, (N) nitrogen, (O) oxygen, (F) fluorine, (Ne) neon, (Na) sodium, (Mg) magnesium, (Al) aluminum, (Si) silicon, (P) phosphorus, (S) sulfur, (Cl) chlorine, and (Ar) argon

2.4.1 Elements and Atoms

Considering what we have learned about gases, it is very surprising that only about one hundred years have passed since the raging debate between Boltzmann and Mach about whether or not atoms exist. It was clear already some time before Boltzmann that our surroundings are composed of a number of so-called "elements." One such element was hydrogen, another oxygen, both known to be gases but also known to occur in liquid and solid form such as in water and ice, respectively. The number of known elements was increasing over time. There were only 40 elements known to the famous French chemist Antoine Lavoisier. As these lines are written, we have 117 known elements with the possibility of creating still (very few) more by involving so-called nuclear chemistry (see radioactivity).

The Russian chemist Dmitri Mendeleev is credited with showing that many elements behave chemically in a very similar fashion to a few others. He created a list of elements in which the columns contain the chemically similar ones. The first 18 elements, listed that way, are shown in Fig. 2.34. More complete lists, called the periodic system of elements, can be found on the Internet.

Hydrogen and oxygen like each other and form water molecules when they interact or, as one says, chemically react. We can deduce from Fig. 2.34 and Mendeleev's rules that lithium and sulfur will also like each other and react similarly. To notice all these chemical similarities took, of course, a lot of time and detailed knowledge. Helium, neon, and argon have in common that they do not readily react with any other element on the list. They are singled out by this distinction and called the noble gases. Their opponents on the left side of the periodic system, hydrogen, lithium, and sodium, on the other hand, readily react with all the other elements on the right side of the table, except, of course, the noble gases. Lithium and sodium are metals and conduct electricity. Hydrogen is a gas that

normally does not conduct electricity but also becomes a metal under extremely high pressure and is thus similar to lithium and sodium. Chemical compounds of the first and the seventh column are salts.

The salt that we use for cooking consists of the two elements Na and Cl. These stick closely together and form a "molecule," a new entity that behaves chemically different than its constituents. The attraction of Na and Cl arises from the fact that the sodium does not "hang on" to one of its electrons, while the chlorine not only likes to hang on to its own electrons but also takes easily an additional one. Thus the Na atom loses its electron to the chlorine and becomes positively charged, a so-called positive ion denoted by Na^+, while the chlorine becomes a negative ion denoted by Cl^-. Positive and negative charges attract each other as we know from Sect. 2.2.1 and therefore stick together. This type of bonding of atoms is called ionic bonding.

It turns out that the formation of molecules due to attractive electrical forces is a general feature that is the basis for all the chemical bonding of atoms that form molecules. The atoms on the left side of the periodic system tend to give electrons away and those on the right side tend to accept electrons. However, the degree of electron donation and acceptance varies significantly depending on the participating atoms. The forces of attraction can therefore range from reasonably strong, as for NaCl, to much weaker as for the van der Waals forces discussed in Sect. 2.3.4. The atoms in the middle of the periodic system, C (carbon) and Si (silicon) in Fig. 2.34, have a very special status. They like to donate electrons as much as they like to accept them. In other words they tend to share electrons in various ways and thereby form so-called "covalent" bonds with other atoms. These other atoms could also be carbon or silicon. For example, coal consists largely of carbon, and we are used to associate the word carbon with coal. However, under high pressure carbon atoms form a special (called "covalent") bond to each other. Indeed chemists have been able to fabricate diamonds out of pieces of coal by using very high pressure equipment. Carbon also forms bonds with other carbons and hydrogen, oxygen and nitrogen. Molecules of that kind are the nucleic acids and proteins that make up our body. It is, of course, the goal of chemistry to understand the details of the forces that hold all of these atoms together and lead to the formation of so many different substances. Chemistry can then use this understanding to produce molecules that are useful for humans in a great variety of ways, ranging from the salt in the kitchen to medicines for diseases. This is the reason for the great importance of the periodic system of elements and its understanding. A rough understanding can be gained by the electrical forces that we have just discussed and by the addition of a few rules. These rules deal with integer numbers of electrons that are preferably exchanged and shared between the various atoms of which the molecules consist. Modern quantum mechanics gives an explanation of these rules, and we will give a brief description of this explanation below and a more detailed one in Chap. 5.

2.4.2 Composition of Atoms and Rules for Molecule Formation

We have explained in Sect. 2.3.2 that the physical properties of gases follow from their atomic structure. There is a very large number (see Avogadro's number) of atoms in any larger volume (such as one liter) of a gas. The atoms themselves are very small, and it took, in the past, difficult experiments to determine their size. Nowadays we can make atoms visible with atomic force microscopes as described in Sect. 3.4. "Seen" with such a microscope, the atoms look like small spheres with a typical diameter of the order of 0.1 nm (10^{-10} m). The air that surrounds us is, of course, such a gas consisting of atoms. However, air appears to us like something continuous. For example, when wind blows in our face, or if we wave around with our hands, the air is everywhere. From such experience we could not possibly conclude that 10^{23} atoms of oxygen and nitrogen are impinging on our face or our hands and we could not possibly conclude that the space around us is almost empty except for the very small atoms. Yet, this is exactly how it is as we can see by use of powerful microscopes.

Thousands of years ago, when Democrit (Democritos in Greek) talked about atoms, he did have the idea that the matter surrounding us is composed of small particles that cannot be further divided. He had the right ideas of how things might be, but there was no way to prove any of his ideas. It took more than 2,000 years to find out that atoms consist again of much smaller entities called electrons, protons, and neutrons. We know now that the diameter of a proton or neutron is of the order of 10^{-15} m, and that of the electron is even smaller. One can, of course, not strictly speak of a diameter of an electron, because no microscope exists that would show us some sphere corresponding to an electron, neutron, or proton. We also know now, from detailed measurements and theory, that the neutron and proton have structure and consist of still much smaller entities,the quarks and gluons described in Sect. 5.4.

To understand the chemistry of atoms, it is most important to understand the interactions of electrons and protons and their consequences. Electrons are negatively charged and swirl around the positively charged protons with enormous speed and frequency of oscillations and turns. As we will see when we discuss quantum mechanics in more detail, no distinct pathways of the electrons can be determined. However, there are patterns to this swirling of the electrons that can be compared to the patterns of standing waves of strings and other objects as we have discussed them in Sect. 2.1.5. The patterns are, in general, three dimensional meaning that if we introduce coordinates (x, y, z), all three directions are of importance. Several forms of these patterns are shown in Fig. 2.35.

Protons and neutrons have a mass that is about 2,000 times the mass of the electrons, and are located in a very small volume, around the zero point of the coordinate system Fig. 2.35. The collection of neutrons and protons at the center of the atom is called the nucleus of the atom. Some of the electrons of atoms swirl around the nucleus in s-patterns that have the form of a sphere, as shown at the top of Fig. 2.35. The electrons can be anywhere inside the sphere and even outside the

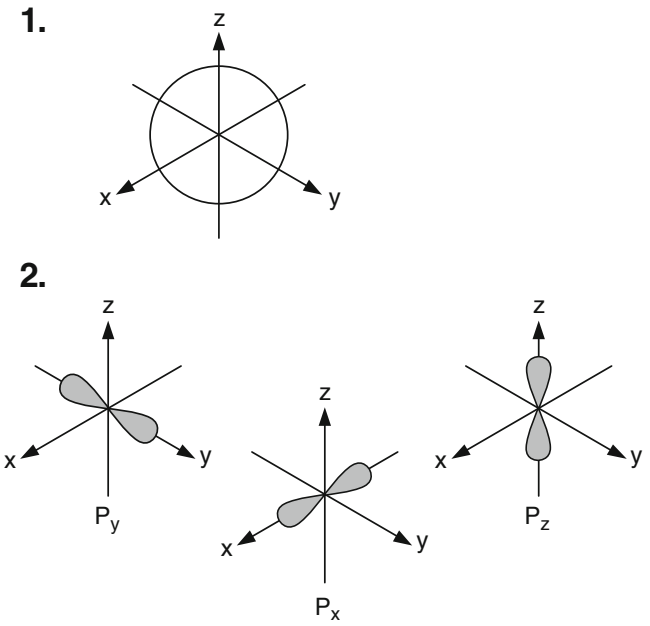


Fig. 2.35 Patterns of electrons "swirling" around the atomic nuclei that consist of protons and neutrons and are located in the center of the shown coordinate system. The *top* pattern (1) shows a circle that symbolizes the symmetry of a sphere and is called the s-type standing wave pattern or s-state of an electron. The probability of finding an electron is largest at the center of the sphere and diminishes rapidly away from the center. The *bottom* (2) shows the more complicated so-called p-type standing wave patterns. They have the form of the ∞ symbol rotated around the axis that is shown. The probability of finding an electron is again largest at the center of the shaded volumes and diminishes rapidly away from the axis around which the shaded volumes are wrapped. Other even more complicated swirling patterns (d, e, f) are also possible

sphere, and they propagate through the nucleus without any interaction other then the electrical attraction. The electrons also have a characteristic energy. We denote this energy here just by an integer number. For example, 1s means the electron exists in an s-type standing wave pattern and has the lowest energy. 2s means the electron is also in an s-type standing wave pattern but now at a higher energy with the label 2. We can also have 3s, 4s, etc. It is important to note, that the electron cannot get "stuck" in the nucleus because then it would loose its kinetic energy and such processes are not possible and energy must be conserved. Therefore, the electron must just continue in its swirling pattern forming a stable atom while moving and moving. There are also other patterns that electrons follow around the atoms. Three such patterns of the so-called p-type are also shown in Fig. 2.35. Again these patterns can correspond to different energies, and we then write 1p, 2p, 3p,. . . etc. There exist more complicated patterns that are important for the atoms beyond number 20, but we will not discuss them here because we wish to discuss only the major principles.

Two important rules that we need to remember are the following:

1. Each atom has an equal number of electrons and protons. This number equals the number given to the atom in the periodic system (hydrogen has the number 1, helium 2, lithium 3, etc.).
2. Each standing wave pattern with a given energy (labeled 1, 2, 3,...) can accommodate exactly two electrons. Energy 1 (the lowest energy) has only s-pattern standing waves, energies 2 and 3 have the possibility of both s- and p- pattern standing waves. The fact that each pattern with given energy can accommodate precisely 2 electrons, follows from the Pauli Principle named after the Austrian physicist Wolfgang Ernst Pauli and is related to the so-called "spin" of electrons discussed in Sect. 5.3.2.

Using these rules, we can find the electron patterns of the atoms of the periodic system of elements. We assume that the electrons will be in their lowest form of energy which is normally the case. Hydrogen (H) has only one electron. Therefore the hydrogen electron has a 1s pattern. Helium has two electrons and therefore two 1s pattern electrons according to rule (2). Because helium is a noble gas, this means that two electrons in the 1s pattern form a very stable and nonreactive atom. Lithium has two 1s electrons and one 1p electron. We note that lithium is very "willing" to give away the 1p electron because then it is left with the two 1s electrons exactly like a noble gas. We make now a jump and go to Neon. Neon has 10 electrons and therefore two electrons of each of the patterns $1s, 2s, 2p_x, 2p_y, 2p_z$. Neon is also a noble gas, and the pattern of Neon must therefore be a stable and unreactive one. Sodium (Na) has 11 electrons and therefore two electrons of the patterns $1s, 2s, 2p_x, 2p_y, 2p_z$ and one 3s electron. Again, in chemical reactions, sodium likes to get rid of the one 3s electron and then to assume the pattern of the noble gas neon. We make another jump to chlorine. Cl that has 17 electrons, two of each of the patterns $1s, 2s, 2p_x, 2p_y, 2p_z, 3s, 3p_x, 3p_y$ but only one more electron of the p_z pattern type because Cl has 17 electrons. The next higher noble gas is argon (Ar) with 18 electrons and therefore two electrons of each of the patterns $1s, 2s, 2p_x, 2p_y, 2p_z, 3s, 3p_x, 3p_y, 3p_z$ which again is a most stable and unreactive configuration. Therefore when compounds of chlorine are formed, Cl likes to obtain the one electron to approach the configuration of argon with 18 electrons.

We can now see how these rules explain the formation of the NaCl molecule. The sodium donates its one electron to the chlorine and the positively charged sodium has then the configuration of Neon, while the chlorine takes an electron, becomes negatively charged, and ends up with the configuration of argon. Thus we end up with a molecule of atomic constituents that do not like to react chemically anymore, because they are similar to two noble gases. In addition, the two atoms of the molecule stick together, because of the electrostatic attraction of positive and negative charges. This type of bonding is known as ionic bonding. Molecule formation can be generally explained based on the basis of these principles, even if the molecules are composed of many atoms. There are, however, complications when it comes to how the electrons are actually shared by the atoms. The donation of the electron to the other atom is in a way the simplest form of sharing. However, this

type works only for the elements from the first and the seventh column. The sharing between two or more carbon atoms, that form a so-called covalent bond, can be very complicated. It took many decades of chemical research to find out how this sharing works in all the interesting chemical materials, starting from salts like NaCl and going to DNA molecules with millions of atoms. In the following, we discuss important molecules and explain how one can understand their formation from the properties of the constituent atoms and their place in the periodic system. We also introduce a shorthand way of writing or drawing the chemical composition of molecules that permits us to see how electrons are shared.

The way to write or draw the chemical composition of molecules is based on the three following facts: (a) the electrons of atoms form s- and p- type standing wave patterns corresponding to energies 1, 2, 3. . ., (b) the s- and p- standing wave patterns can accommodate each 2 electrons, and (c) the electrons of the various atoms are shared in molecules in such a way that the electron numbers of noble gas atoms are approached. For all molecules that we discuss here, this means that we will have constituent atoms with either two s-electrons corresponding to the lowest energy labeled by 1 (as in helium) or eight electrons, two s- and 6 p-type, as in neon (energy 2) or argon (energy 3).

The sharing of electrons by the atoms of the molecule is usually indicated by a line that represents two electrons. The line is drawn between the atoms that share the electrons. Sometimes electrons are simply represented by one dot for each electron. Expert chemists use lines for shared electron pairs and dots for the unshared electrons. The corresponding symbols are called Lewis structures after the chemist Gilbert N. Lewis. We do not emphasize such details of meaning when discussing chemical symbols here, because we just like to drive home the fact that, when molecules are formed, atoms attempt to achieve the noble gas configuration (configurations with eight electrons for all cases that we consider). If the sharing of electrons is important for the explanations, we will emphasize this in the text. Here are a few examples.

A gas of atomic hydrogen H is not stable because the hydrogen atoms like to form pairs with two shared electrons, one from each atom. The corresponding symbols for the hydrogen molecule are H–H or just H_2. One line corresponding to two electrons is called a single bond. The hydrogen molecule has thus two shared electrons for each hydrogen atom and, therefore, resembles the very stable helium atom (2 electrons of the 1s type) that does not tend to undergo further chemical reactions.

The air we breath consist mostly of oxygen O and nitrogen N. Atomic oxygen is also not stable and forms the pairs O = O or just O_2 that are now held together by so-called double bonds, 4 shared electrons of energy 2. Each oxygen atom of the molecule has, in addition, 4 more unshared electrons of energy 2 that are often not indicated (symbol O = O), but could be by writing the symbol as ::O = O::. Together, the total number of unshared and shared electrons of energy 2 is thus 8, exactly the number of the noble gas neon. Nitrogen forms molecules with three electron pairs shared by each atom plus two unshared electrons, resulting in N ≡ N or : N ≡ N : or just N_2. In this way, we have again the 8 electrons of energy 2 for each nitrogen atom, just as many electrons with energy 2 as neon has.

As we know already, for the combination of Na and Cl, we have no sharing of electrons; the electron simply moves from the sodium to the chlorine to form positively charged Na^+ and negatively charged Cl^- ions (each resembling a noble gas). Na^+ and Cl^- stick together by the electrical attraction and form NaCl molecules.

General molecule formation is a very complex problem and its corresponding chemistry requires significant expertise. For example, the electrons of oxygen can be shared in a different way then for O_2, and the earth's atmosphere contains small quantities of O_3, also known as ozone. The interested reader is referred to descriptions on the Internet. With more than three atoms and different types of sharing or nonsharing (as in NaCl), one deals with an enormous amount of different ways to form molecules. The formation, fabrication, and use of these molecules are the essence of the art of chemistry. We are not explaining how to actually fabricate molecules from their chemical constituents, but just introduce a number of molecules and their properties as well as their importance for humans.

2.4.3 Important Molecules: From Salt to DNA

No attempt is made to present a complete list of important molecules; there are simply too many. The molecules chosen are just representative examples to show how relevant chemistry is for our life and are supposed to stimulate interest. The true chemist and student of chemistry needs, of course, access to a lab and needs to perform experiments. There are also chemistry kits available for home use, and the Internet lists many of them.

Molecules in our Household

We have already discussed the chemistry of salt and have seen that the bond between Na and Cl arises from the electrical attraction after the electron is transferred from Na to Cl. This ionic bond is characteristic for all salts, also those that are not used for cooking such as LiF or MgF_2. You can construct all possible salts from the periodic table and from our rules. They always involve a metal such as Li, Na, or Mg on the left side of the periodic table and other atoms such as F or Cl on the right side. The metals like to donate their electrons, and F or Cl is happy to accept the electron, and the ionic bond is formed. The resulting ions must have the configuration of the noble gases, and this rule tells us how many ions we need, for example, two fluorine ions for one magnesium ion or just one sodium for one chlorine. As you may have noticed, we have not included into this discussion molecules that involve hydrogen such as HCl. Indeed such a molecule has features that are very similar to NaCl, and the attentive student may ask the question why we have not listed HCl as another salt? Indeed HCl is a very important molecule. However, there is a big difference between a sodium atom and a hydrogen atom that both have lost one

electron: the sodium ion (atom minus one electron) has a noble gas configuration while the hydrogen ion (hydrogen minus one electron) is just a proton with no electron at all. This fact has the following consequences.

HCl molecules form a gas that can be dissolved in water. This solution is a strong electrolyte, meaning that the H and the Cl are completely ionized. In contrast to a solution of NaCl, however, the HCl solution is not just a salty liquid but is very reactive and attacks many substances. It is called a strong acid. Why is HCl much more reactive in certain ways than NaCl? The reason is the following. Hydrogen does indeed donate its electron and the chlorine takes it, just as we had it for Na and Cl. However, once hydrogen has donated its electron, there is no electron left, just the positively charged proton. The positively charged Na^+ owns still electrons and thus resembles the noble gas Ne. The proton does not compare to any noble gas. It actually does like to react with other atoms, unlike the noble gases do and therefore protons lead to acidity. In fact acids are just molecular entities that can supply protons. It is interesting that the name proton is of rather recent origin; it was introduced by Rutherford in 1919 during his investigations of atomic nuclei. Rutherford's discussion and naming of protons was seen at the time, and still appears to some, remote from the experiences of daily life. Yet, humans have tasted protons since ages in their drinks. There exist great varieties of acids. Weak acids are in many of our drinks such as lemonade or Coca Cola. Watery solutions of HCl form strong acids. Nevertheless HCl resides in our stomach. It helps turning solid food into liquid and starts the digestion process. HCl also destroys bacteria and therefore protects us from illness. Nevertheless, sometimes our stomachs contain too much acid and then the stomach walls are attacked and suffer, leading to stomach pain and ulcers. The medicines that are then prescribed are proton inhibitors, substances that inhibit the supply and availability of protons in our stomach.

The acidity of food needs to be controlled very carefully. For example, if we buy chicken soup, it has to be very close to "neutral" and not acidic at all. Food science and engineering require, therefore, a very careful control and measurement of acidity. The acidity has its own scale like temperature and is measured by so-called pH values. The symbol pH comes probably from "power of hydrogen." Pure water is said to be neutral (not acidic or tasting sour) and is assigned a pH value of 7. A lower number means that the substance is acidic, 1 being very acidic. The pH scale is a logarithmic scale which means in essence that every point moves you a factor of 10. A pH value of $\mathrm{pH} = 1$ means that the substance is 10 times as acidic as $\mathrm{pH} = 2$ and 100 times as acidic as $\mathrm{pH} = 3$. A pH value above 7 means that the substance does not like to provide protons but instead likes to accept protons. Such substances are called bases. An example for a strong base is the molecule Na^+OH^- because it likes to accept the proton in the chemical reaction

$$Na^+OH^- + HCl \rightarrow Na^+Cl^- + H_2O; \tag{2.71}$$

in words, the base NaOH added by the acid HCl results in salt and water. Salt is still a little corrosive but not as bad as the acids or bases are. If the simple ionic bond can already have so many consequences for the science of our food, imagine what

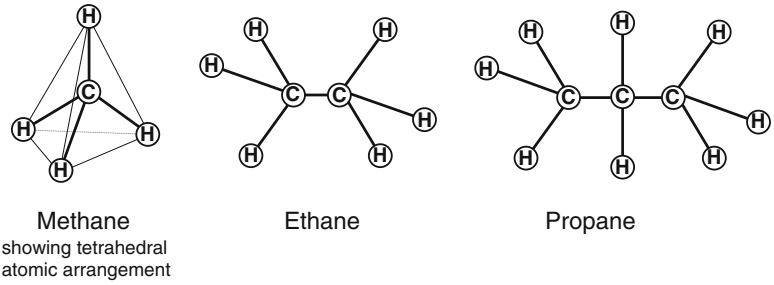


Fig. 2.36 Three so-called hydrocarbon molecules consisting exclusively of hydrogen and carbon. The nearest neighbor atoms of each molecule are arranged in the geometrical form of a *tetrahedron*. The illustration attempts to show this 3-dimensional arrangement

effect all the other forms of chemical bonds have, especially covalent bonds and the participation of a multitude of atoms.

Of particular importance are the following carbon-related molecules. Carbon (C) is a very special element because it is the first element that is located in the center (4th) column of the periodic system. It has therefore altogether 6 electrons, two each of energy 1 s-type and energy 2 s-type as well as energy 2 p-type. Thus it has 4 energy 2 electrons and needs another 4 energy 2 electrons to achieve the chemically inactive neon configuration. There are many different ways of receiving these 4 additional electrons by sharing with other atoms. The most straightforward way for carbon to get 4 additional electrons is to share electrons with 4 hydrogen atoms. CH_4 molecules are the molecules of the gas methane. The next more complicated way of sharing is to have two carbon atoms that share two electrons with each other, and each of the two carbons shares electron pairs with three hydrogen atoms. C_2H_6 is called ethane, and one can go on like this and find molecules of gases and liquids that all burn and combust well and are therefore of use for gas grills or to drive combustion engines. Gasoline is made of such molecules that are called hydrocarbons. Figure 2.36 shows the first three hydrocarbons. Note that each carbon is "connected" to 4 lines, with each line (or two dots) representing a pair of shared electrons and thus altogether 8 electrons.

The true shape of the molecules is, of course, three dimensional, and some attempt is made in Fig. 2.36 to indicate the three-dimensional arrangement of the atoms. It is important to remember that carbon likes to arrange its neighbors in the form of a tetrahedron: the four lines of shared electron pairs point to the corners of a tetrahedron as indicated in the illustration. In general, it is difficult to plot molecules in three dimensions. Some texts for chemistry try to introduce certain ways of shading the atoms to indicate in which direction they are pointing. We have not attempted to include any special way of 3-dimensional rendering. Many Internet sites have moving or rotating pictures of important molecules so that you can get a better 3-dimensional feel for them.

Adding different kinds of atoms such as oxygen (O) brings us to another important type of hydrocarbons. We know that the atom combination –O–H

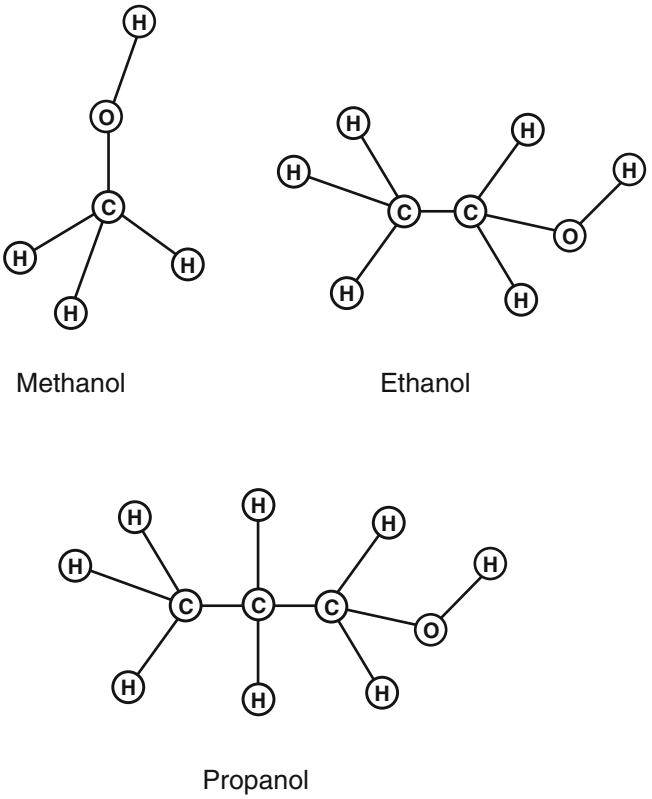


Fig. 2.37 Alcohol molecules that are obtained from the corresponding hydrocarbons by replacing one H atom by -O-H

"desires" to share an additional electron. Adding hydrogen that contributes one electron for the shared pair, one obtains the water molecule H–O–H or H_2O. We can also insert the –O–H group instead of one of the hydrogens in the molecules of Fig. 2.36 to arrive at new molecules that are shown in Fig. 2.37. These molecules are called alcohols. The second, ethanol, is especially well known as a component of drinks such as wine and liquors. It is also combustible and is often used as fuel for racing cars. Ethanol, particularly mixed with gasoline, is also frequently used as fuel for ordinary cars.

The molecules described above can be constructed from rather simple considerations of electron sharing. As we will see in the section on quantum mechanics, this sharing means that the electrons form something like a standing wave, an analog to the vibrational patterns of objects such as strings or drums, in between and around the various atomic nuclei. For single atoms, these patterns are, as we have learned, of s- and p-type (or d-, e-, f-type for the elements with higher numbers than those listed in Fig. 2.34). The molecules discussed above can still be explained in terms of these patterns and the tendency to form configurations that are close to that of

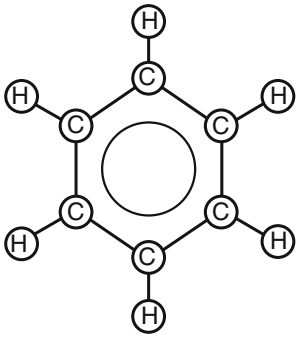


Fig. 2.38 Benzene molecule exhibiting a ring-like structure. Every *straight line* of the illustration represents the sharing of two electrons. The *inner circle* indicates six electrons that are shared by all carbon atoms and move freely between them. This special sharing of many electrons makes the molecule stable, because in this way the carbon atoms approach the noble gas electron configuration

the noble gases by sharing pairs of electrons. However, there exist many molecules for which the sharing is of different nature. An important example is benzene, a smelly liquid. The molecule of benzene has 6 carbon and 6 hydrogen atoms, and its chemical symbol is therefore C_6H_6. How on earth can the electrons be shared to approach the noble gas configuration (eight electrons of energy 2 for carbon and two of energy 1 for all hydrogens)? Well, there is a way, and it is shown in Fig. 2.38.

The first new fact that we notice is that the molecule contains a closed ring of carbon atoms. This closed ring provides, of course, new possibilities of electron sharing. We could, for example, have two hydrogen atoms for each carbon leading to a molecule C_6H_{12}. Such a molecule does indeed exist and is called cycloalkene. The benzene molecule, however, has another novelty in it. This novelty is the sharing of carbon electrons among all the carbons. We can see that ring-like structures will therefore add numerous possibilities to molecule formation. The number of possibilities becomes even larger, if we admit atoms other than carbon into the ring itself. The possibilities of sharing electrons between all of these atoms increase significantly, and the understanding of this electron sharing between many atoms including rings becomes a real art. Chemistry is therefore divided in subdisciplines such as inorganic chemistry (the chemistry of NaCl, HCl, etc.) and organic chemistry which is in essence the chemistry of carbon-related molecules. Biochemistry is related to the molecules of living beings that are carbon-related.

We finish this section by mentioning a molecule of greatest importance: glucose. The name derives from the Greek word “glukus” for sweet. Glucose is a form of sugar that our cells use as a source of energy. It exists as a string of carbons with oxygen and hydrogen atoms attached resulting in the molecule $C_6H_{12}O_6$. There exist also several variations of glucose that contain a closed ring in which one of the carbon atoms is replaced by oxygen.

Sugars are of far broader use and presence than just for the energy supply of many life-forms including humans. Sugar molecules can be put together in long

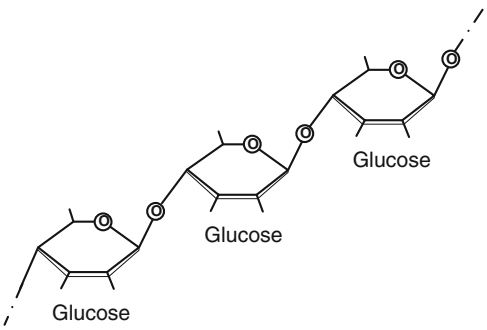


Fig. 2.39 Schematic of a single strand of cellulose, a very long molecule. Many such intertwined strands form the basic material for plants

chains that consist of up to 10,000 glucose molecules connected by oxygen. These chains then form cellulose, a very important basic material. The long molecules intertwine with other long molecules to form the stable material of which green plants are made. Such a material made of intertwined chains is called a polymer. Poly is the Greek word for many. Plastic materials are polymers and are very important for our daily life. Figure 2.39 shows a single strand of cellulose.

You can imagine the enormous possibilities of constructing and using molecules, be they chain-like or ring-like and polymers of them that form three-dimensional structures and materials of all kinds. This is naturally great fun for the real chemist. Chemists analyze molecules and also put them together (synthesize). They have synthesized enormous numbers of useful molecules including the most significant and important of all: DNA. DNA is what we discuss next.

Molecules We Are Made of and Molecules that "Make" Us

The molecule that is central to life, including human life, is the DNA molecule. This is a giant, long molecule like cellulose. Remember that cellulose consists of smaller sections of sugar-like molecules that are repeated and repeated all over and connected by chemical bonds. DNA has a "backbone" or "skeleton" (also "spine") made out of repeated sugar-phosphate units (see Internet for the phosphate molecules) to form long strings of repeated molecule sections. However, the really important ingredients of the DNA are chemicals that are denoted by the letters A (adenine), T (thymine), C (cytosine), and G (guanine). Early chemical analyses by Erwin Chargaff had shown that each giant DNA molecule contains equal quantities of adenine and thymine and also equal amounts of guanine and cytosine. This was an important clue for the two young scientists, James Watson and Francis Crick, who came up with the precise structure of the DNA in 1953. What does DNA actually do and how does this molecule look like?

DNA contains a code that dictates how proteins, the materials that make up our bodies, are produced. Proteins, in turn, determine the general properties and functionality of our bodies, such as muscular strength. The code of the DNA is

similar to any code that can be expressed by a few symbols, such as the binary codes that form computer programs and are based on the symbols (bits) 0 and 1. In the case of DNA, we deal with the four symbols A, T, C, and G that stand for the corresponding molecules. A message of the DNA is therefore equivalent to a string of these symbols, as for example

$$\text{ATCGATTGAGCTCTAGCG.} \tag{2.72}$$

According to Chargaff's rule, one is forced to assume that the DNA that contains the above string will also contain a second string of code given by

$$\text{TAGCTAACTCGAGATCGC,} \tag{2.73}$$

because we need to come up with equal numbers of (A, T) as well as (G, C), respectively. Thus, the code of the DNA must occur in two strings, the second of the two being obtained from the first by exchanging adenine (A) with thymine (T) and guanine (G) with cytosine (C). Only in this way can we end up with a given arbitrary code and equal quantities of adenine and thymine as well as guanine and cytosine. The two lines above look like the "positive" and the "negative" of film. At the time when film was still very important, the words positive and negative reminded everyone immediately of making copies of photos. Nowadays, with digital cameras that work without paper, different thoughts are connected to the processes of making copies. Nevertheless, it was an important clue for the discoverers of the DNA structure that the process of copying DNA molecules that we describe below, was somehow already contained in the structure of DNA and could be, and as we now know is, the basis of the reproduction of life. The actual DNA molecule is indeed made of two strands of code, like the ones shown above only much longer. These two strands are mounted on a sugar-phosphate skeleton, and the skeleton and strands are twisted to form the so-called "double helix" that is shown in Fig. 2.40.

The word double helix indicates that there are two spirals nested in each other. The outer boundary of the spirals is formed by the sugar-phosphate skeleton, and the spirals are bound to each other, because the A, T molecules and the G, C molecules match up and stick together. This sticking together is accomplished by hydrogen-sharing bonds, as we have discussed them in connection with van der Waals forces and the sticking together of water molecules that was illustrated in Fig. 2.31. The A, T, G, C molecules contain rings of carbon with nitrogen and oxygen substitutions. The hydrogen bridges are located between the nitrogen atoms or between nitrogen and oxygen atoms and are shown together with the molecules in Fig. 2.41.

Any given cell of any living being, for example, a bacterium or a human being, contains at its core such DNA molecules. The number of A, T and G, C pairs, that a cell contains, depends on the sophistication of that being. For humans, the number of pairs is about 3 billion. The first artificial DNA, as synthesized by G. Craig Venter and colleagues, was for bacteria and contained "only" about a million pairs. In spite of the rather small number of (A, T) and (G, C) pairs of this artificial DNA, the experiments of the Venter group have a great significance. The artificial DNA was introduced into lifeless bacterial cells that did not contain DNA. These cells started

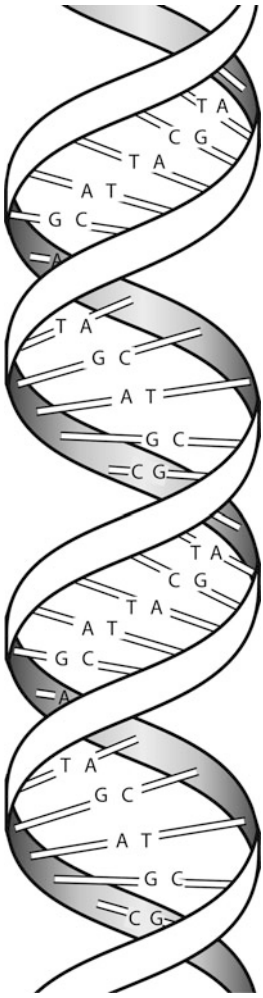


Fig. 2.40 A section of a DNA molecule. The winding "tapes" symbolize the sugar-phosphate skeleton. A code of molecular sequences of (A, T) and (G, C) pairs is also shown. Notice that we are attempting to show a 3-dimensional molecule. Shorter molecule "connections," as, for example, the =T A= at the *top* of the figure, indicate only that the connection points out of the page, while the longer connections are parallel to the page

then to divide and multiply just as natural living cells do. This means that the science of DNA is a step closer to a predictive and experimental repeatable understanding of bacterial life-forms. The process of DNA division and replication and multiplication is very similar for all life-forms and works as follows.

In the first step the DNA must be unwound and "unzipped" into two strands, a "positive" and a "negative," as discussed above. This is accomplished by the presence of a chemical (enzyme) called helicase. When this chemical is supplied, the cell starts to divide. Then the hydrogen bonds are broken and the two strands of the double helix are pulled apart forming two single strands. In the next step these two strands are immersed into a different "soup" that contains another enzyme called DNA polymerase as well as a sufficient supply of A, T, G, and C molecules, and each molecule of each strand will pick up its exact partner. On the other side

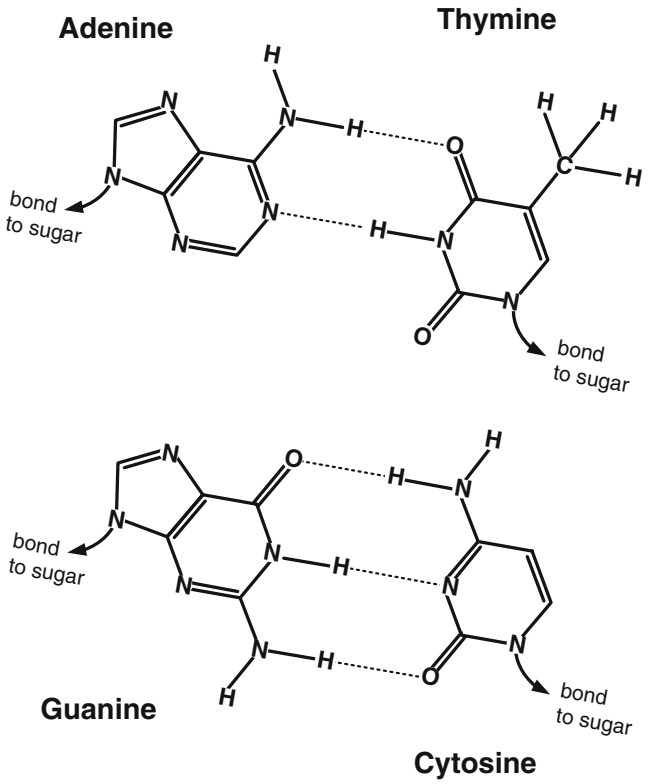


Fig. 2.41 The molecules adenine, thymine, guanine, and cytosine contain rings that include carbon (not indicated but located at *ring corners*) and nitrogen (indicated by N). Adenine and thymine are connected by two hydrogen bonds, guanine, and cytosine by three. The bonding to the sugar-phosphate skeleton (not shown) is indicated by *arrows*

of these new partners, the sugars and phosphates join together and form the new skeleton, and now you have two new DNA double helix strands. The double helix strands are the core material of the so-called genes that contain all information that makes our cells "tick," the information and properties that we inherit from our parents, and that we hand down to our offspring.

Of course, something needs to be done with the information that the DNA provides. Some materials need to be formed, in order to end up with new beings, be they bacteria or humans. This process of transforming the code of the DNA into actual materials is by now reasonably well understood. The materials, that are built by the coded information, are the so-called proteins. Proteins are the stuff we are mostly made of; they form our organs and muscles, and are polymer materials (like cellulose, the basic materials for plants). Proteins are composed of only 20 different chemicals that we either obtain from food or that our body produces. These 20 different chemicals are called amino acids. The formation of any given amino acid

takes triplet DNA code sequences, such as CAA or CAG, that are called codons. Each of the elements of a triplet may be one of the 4 basis A, C, G, and T. Therefore, there exist $4 \cdot 4 \cdot 4 = 64$ different code triplets that lead to the production of the amino acids. Because we have only 20 amino acids and 64 codons, the coding is redundant. For example, both CAA and CAG lead to the production of the amino acid called glutamine. A string of 300 such codons instructs the cell to build a protein consisting of 300 specially arranged amino acids. All the complicated proteins, that our body needs and consists of, can be generated by codon strings. It is interesting to note that almost all life-forms, from algae to humans, use the same codons, hinting that all these life-forms have developed in similar ways. This fact is one of the cornerstones of the theory of evolution teaching the evolution of more complex life-forms from simple ones.

The details of this story, how cells multiply, how information is transmitted from the parents to offspring, and how all the materials that are involved are coded and produced, all of these, are currently explored in highly interesting and complex research projects. These projects are relevant to everything that concerns us humans, from health to illness and to hereditary traits. The projects encompass the composition of the materials that make our bodies and extend to complicated personality properties such as a hot temper. There is hope that all of these important problems related to how we live and die will one day be understood as well as Euclidean geometry is understood now. This is a wonderful playground for those interested in STEM and, particularly of course, those interested in biochemistry.

Crystals: Metals, Semiconductors, and Insulators

Cellulose and DNA are giant molecules. Many intertwined cellulose strands form leaves and stems of plants and can form all kinds of 3-dimensional shapes. The DNA in the human genes contains billions of A, T, G, and C molecules and would be several inches long if stretched out. Characteristic of both types of giant molecules is a lack of regularity. This lack of regular arrangements of atoms arises, in the case of cellulose, from the intertwining of the strands that is typical for polymers. Polymers are, thus, not very "regular." In the case of the DNA, it is the codes that can be all different, and the lack of regularity is grounded in the very nature of DNA and its complicated coding system. We can compare this lack of regularity with that of books: there is some regularity in the arrangement of the letters and words, but the text is naturally all different and describes different things. Crystals, on the other hand, are giant collections of atoms (or molecules) with a completely regular arrangement of the atoms (or molecules). The Internet teaches you how you can easily grow salt crystals of centimeter (0.01 m) size. Such a crystal contains more than 10^{23} Na and Cl atoms that are all arranged in a regular pattern as shown in Fig. 2.42. The NaCl crystal is, electrically speaking, an insulator, meaning that it does not conduct electrical currents. The Na^+ and Cl^- ions are fixed in the crystal and cannot be moved by normal electrical forces such as available from a battery.

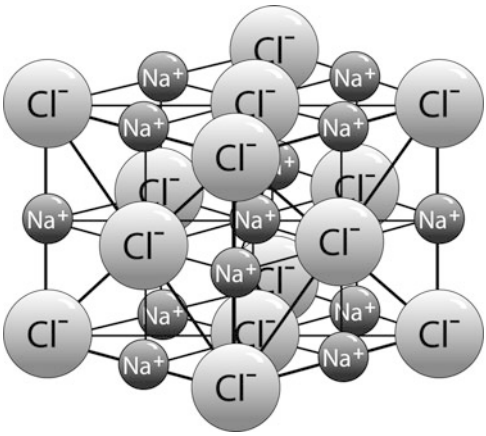


Fig. 2.42 Geometrical arrangement of Na^+ and Cl^- ions forming an NaCl crystal. Notice that the arrangement of the atoms is cube-like as indicated by the lines. This cubic shape is also maintained for very large crystals. The atomic configuration of a large crystal is obtained by simply continuing the shown pattern to all sides in such a way that any given "face" of the cube becomes the first layer and face of the next cube. All such cubes then flawlessly assemble to form the big crystal

The electrons of the two ions are also localized with these ions and cannot easily move to neighboring ions. Ordinary electric fields are simply to weak to move the electrons from one such ion to the other.

Sodium (Na) by itself can also form a crystal, and such a crystal has also the symmetry of a cube. However, the sodium crystal is a metal and an excellent conductor of electricity. The reason is simply that sodium likes to give away one electron when connecting with other atoms, in order to achieve the neon configuration. Therefore, all Na atoms of the crystal give away and share the outermost (energy 3 and s-type) electrons of the single sodium atoms. Because these electrons are shared by all atoms, they can freely move in between them, and thus sodium crystals conduct electricity. The bonding of the sodium atoms by all the shared electrons is called metallic bonding. This type of bonding is relatively weak and is characteristic for all metals (this is why gold can be formed into jewelry).

Crystals made of elements from the 4th column have different and very special properties. Carbon atoms can form a variety of crystals, depending how they are arranged and share electrons. Graphite is one of these carbon crystal types and conducts electricity as metals do. Graphite consists of sheets (thin two-dimensional layers) of graphene that sit on top of each other and are only loosely held together by weak bonding forces. Graphite is relatively soft, because the two-dimensional layers can slide on top of each other. They also can be pealed off, and graphite is, therefore, used as the writing material in pencils. The single graphene layer sheets themselves, however, are much more robust and show that the graphene bond is much stronger than that of metals and is based on a special sharing of the electrons. The best known carbon crystal, with very strong bonding, based on another similar type of special electron sharing, is diamond.

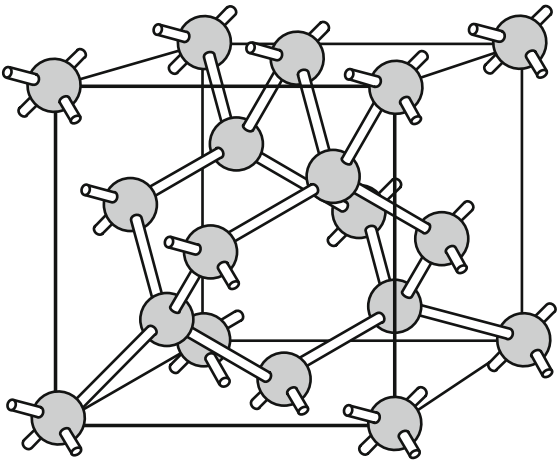


Fig. 2.43 Arrangement of carbon atoms in a diamond crystal. Notice the tetrahedral arrangement of the nearest neighbors and the resulting shape of a cube. A large crystal is obtained, as in the case of NaCl, by continuing the pattern to all sides

Diamond is known to be a very tough substance that cuts glass, and pure diamond is an insulator, that is, it does not conduct electricity. This fact indicates that the bonding of carbon atoms in diamond is not based on the metallic sharing of electrons. The carbon atoms of diamond are located in a tetrahedral arrangement of the nearest neighbor atoms as shown in Fig. 2.43. As also shown in the figure, the atoms still fit in the form of a cube with every face of the cube containing atoms at each corner and one in the center. Such a crystal lattice (the set of all dots that symbolize the atoms) is called a cubic face-centered crystal lattice. The bond between the carbon atoms is called a covalent bond, meaning that basically only neighbors share these electrons.

The lack of electrical conduction in diamond crystals is only observed for "pure" crystals. If we introduce into the diamond small amounts of atoms of a different kind, such as the right-hand neighbors (in the periodic system) nitrogen (N) or phosphorus (P), then diamond does conduct electricity. The reason is simply that N and P have one more electron compared to carbon, and they like to give away this electron to the carbon atoms. This additional electron can then be shared by all carbon atoms and can therefore contribute an electric current. Such additional atoms are called "donor atoms," because they donate electrons to the insulator and make a conductor out of it. One cannot put arbitrary amounts of donor atoms into a crystal, because this would either ruin the crystal or lead to the formation of a different crystal. Usually one can only introduce as much as about 1 % of the donor atoms relative to the crystal atoms. Therefore the conduction of such crystals with donors is not as good as the conduction of gold or the Na crystal that are highly conducting metals. Crystals like diamond, particularly when supplied with donors, are therefore called semiconductors. The important factor for semiconductors is that the conduction can be controlled and therefore "engineered."

There is another very important factor that makes semiconductor materials so important for electrical engineering. This is the possibility of introducing so-called

acceptors such as boron (B) and aluminum (Al) atoms that are to the left of C in the periodic system. Boron has one electron less than carbon. Therefore it tries to take away one electron from the rest of the crystal so that it would be configured similar to the carbon atoms. Because an electron is taken away, a positive "hole" (lack of negative electron) is created in the carbon crystal. Other electrons of the carbon crystal can jump into the hole. Such a jump creates another hole, and therefore positive holes can move and conduct electricity as the shared electrons do. A hole is thus a *missing* electron and is shared by all crystal atoms. It can be regarded as a freely moving positive charge that contributes to conduction. The number of holes can be chemically controlled and engineered.

As we will see in later chapters, the conduction by negative electrons and positive holes is the secret of all of modern electronics. Electrons and holes in semiconductors form the basis of our computer revolution. You might say now: "yes, but are diamonds not too expensive for electronics?" The answer is, of course, yes, they are! In addition, diamond crystals are much more expensive if they are bigger, like the ones in the crowns of kings and queens. Fortunately, there are more inexpensive crystals that can be used for electronics. Silicon is in the same column of the periodic system as diamond and has therefore very similar properties. The electronics industry uses silicon to grow their crystals. Silicon crystals are easier to grow and do not require the high pressure that diamond formation needs. Silicon crystals can be grown with diameters up to around 0.6 m. These are really giant crystals that can then be sliced into thin sheets called wafers. Wafers are used to produce the electronic chips that are in our computers, in cell phones, and in almost all instruments of our households. More of this will be discussed in Sect. 3 on engineering and technology.

2.5 Energy of Atoms, Electrons, and Photons

From the above description of atoms, of molecules, and of the electromagnetic forces, it follows that almost everything that surrounds us, all that we can touch, smell, and see, is either made out of atoms (positive nucleus plus negative electrons) or can be described by electromagnetic waves including visible light. This section describes what we know about the nature of electrons and light and some of the important equations for the energy of these entities in atoms, in molecules, and in free space. The classical description of physical phenomena is usually given in terms of either particles or waves. Billiard balls are particles, and the wireless communication discovered by Hertz is accomplished with electromagnetic waves. As physics evolved into its modern form, it was found that nature does not present us with a clear distinction between particles and waves, certainly not when we deal with the very small, with atoms, electrons and molecules. As science penetrated down to the atomic scale, it was found that the clear distinction between particles and waves is lost and the classical laws need to be replaced by an "amalgamate" of

wave and particle descriptions. This amalgamate is known as the quantum theory of matter and is described in this section in a phenomenological way and in more mathematical detail in Chap. 5.

2.5.1 Light: Waves and Particles

Newton thought of light as particles. Maxwell came to a different conclusion as discussed in Sects. 2.2.4 and 5.3. His wave equation for electromagnetic phenomena definitely suggested that light is an electromagnetic wave. Hertz followed Maxwell's theory and created such waves in the laboratory and demonstrated the possibility of wireless communications. The problem was not settled, however, and investigations of the emission of light from heated and glowing bodies could not be explained by a theory based only on waves.

Photons

Max Planck studied the available data for light emission as a function of temperature and came to the following conclusion that he presented to the German Physical Society:

"We found—and this is the essential point—that the energy E of light must be thought of as being composed of a given number of equal parts. The energy E of these parts can be determined from the equation:

$$E = h\nu, \tag{2.74}$$

where h is a constant given by $h = 6.626\,10^{-34}$ Joule seconds (J s) and ν is the frequency of the light or any other electromagnetic radiation (such as infrared)."

It is pretty amazing how clearly Planck expressed himself. He talked to a room of experts and what he said is, in its essence, understandable to everyone who can read a simple equation. Nevertheless, it was Einstein who made Planck's statement entirely clear. Einstein stated that these "equal parts of energy" are the particles of light. He further postulated that the energy shown in the above equation is the energy of one such light particle. Particles of light are nowadays called photons, and Eq. (2.74) is now called the Einstein–Planck equation. The mathematical-physical thought-process of Planck and Einstein, which took them from the continuous light waves of Maxwell to the postulate that these waves were actually made up of particles (photons), is called quantization. Remember the digital representation of the grey shades and colors of pixels. This representation of shades and colors by binary numbers is also a form of quantization, a quantization that is completely man-made and is introduced only because electronic handling of photos works nicely that way.

The quantization of electromagnetic fields as formulated by Planck and Einstein is, on the other hand, a quantization that we have to introduce, because otherwise we cannot explain the experiments with light. We can see from Eq. (2.74) that these equal parts of energy, the energies of the photons, are extremely small quantities, because Planck's constant h is extremely small. The unit of h is the unit of what one calls the action, which is energy multiplied by time (expressed above in Joule seconds). This complicated unit is needed because the above formula tells us that if we multiply Planck's constant by the frequency ν which has the units $[\frac{1}{s}]$ (one over seconds), then we obtain the energy. The frequency of visible light was discussed in Sect. 2.2.4 and is around $6 \cdot 10^{14}[\frac{1}{s}]$. This means that the visible photons have an energy of about $4 \cdot 10^{-19}$ J, which is a very small number. The energy radiated by the sun onto one square meter of the earth during one second is about 100 J. This means that about $\frac{100}{4 \cdot 10^{-19}} = 2.5\, 10^{20}$ photons hit every square meter of the earth during one second. This is an enormous number. As a homework problem, you can calculate how many photons the whole sun emits during a second. For this calculation you need to imagine a sphere with a radius that equals the average distance of the sun from the earth and you need to calculate the surface area in square meters. It is because of the fact that the numbers of photons is so large that we really do ordinarily not realize that photons behave like particles. This fact can be compared to the case of the pressure of gases that is caused by very many atoms whose single effects and impacts we do not feel at all. We also cannot see or feel single photons. We know, however, that in large numbers, they follow exactly the wave equation and the rules of Maxwell. Therefore photons must be some strange mixture of wave and particle that we called previously already a wavicle. We will see below that such a mixture of wave and particle properties also describes electrons, protons, atoms, and everything else we know. They all behave like waves, yet when we actually measure some quantity, then this gives always the result that the total energy must be thought of as being composed of an integer number of equal parts, just as we would expect for particles.

We know from Sect. 2.1.5 that any wave has a wavelength, the distance between two valleys or two maxima of the wave. Therefore we need to be able to associate a wavelength with electromagnetic waves such as light or radio waves. This is indeed easy to do with radio waves because we can measure the electric field at a number of points in space and what we get is indeed the form of a wave. This type of measurement is already difficult for visible light, because the wavelength is small (of the order of 10^{-6} m) and becomes even more difficult for X-rays that have a very very small wavelength. However, it still can be done because of the phenomenon of diffraction that was explained in Sect. 2.1.5. We have discussed this phenomenon in some detail, because it is at the heart of the wave- and particle-like nature of everything we know, at least as far as I understand it. Diffraction shows that light and all electromagnetic radiation behave like waves do. At the same time we know from Planck and Einstein that such radiation always comes in "lumps" of energy $h\nu$. These facts seem to be in contradiction to each other. But wait, it gets even more interesting when we look at electrons and other "particles."

Wavicles

Diffraction is a general effect for all wavelike phenomena and works for water waves, radio waves, light, and any other wave. The amazing thing is now that diffraction effects can be measured also for electrons, protons, and all other building blocks of our universe, because everything we know is made out of "wavicles"; all constituents of our world behave part time as particles and part time as waves. As strange as this may sound, this is a basic truth of our world, and because it is so basic, we need to discuss it in more detail. Before we do this, we note that the exact calculation of what happens if photons hit gratings (one-, two-, or three-dimensional gratings including crystals) must include all possible pathways the photons can go and not only the two pathways that we have shown in Sect. 2.1.5. This calculation is indeed possible and was pioneered by Richard Feynman. Feynman also showed that this type of calculation works for all wavicles that we know, and he designed a general method that always can be used. This method is called Feynman's path summation or path-integral method. This calculation with all its bothersome details is painfully complicated. Fortunately it can often be replaced by simple rules that one deduces from the addition of a few such photon paths, and there are also modern computer software packages that can deal with it.

2.5.2 Electrons: Particles and Waves

Electrons were at first considered to be particles, in the sense of macroscopic bodies like billiard balls, that had concentrated their mass in a certain volume and also had some charge concentrated essentially in a point. This incorrect, or at least incomplete, analogy received a fatal blow when Louis de Broglie suggested that electrons may also behave like waves, at least partly so. Experiments, that involved crystal materials that can be viewed as natural gratings with a spacing of the lines (atoms) of about 10 nm or less, did confirm the wavelike properties. Electrons were reflected by the crystal gratings exactly as if they had a certain wavelength. This wavelength depends on the momentum $p = mv$ of the electrons, where m is the electrons mass and v its velocity. Louis de Broglie suggested the following equation:

$$\lambda_{\mathrm{dB}} = \frac{h}{p}, \tag{2.75}$$

where h is Planck's constant. This means that we should imagine that the electron is something like a very small particle-like "blob" that behaves also like a wave. The size of the blob is about a few wavelength λ_{dB}. If we wish to calculate that wavelength, we need to know the velocity of the electron, which may, of course, greatly vary depending on circumstance. As is known now from many experiments, the electron does not really stand still under any practical circumstance. In fact, in atoms, the electron velocity is of the order of $3\,10^6\,\frac{\mathrm{m}}{\mathrm{s}}$, which results in a momentum

of $2.73\,10^{-24}\,\frac{\mathrm{kg\,m}}{\mathrm{s}}$ because the mass of the electron is $9.1\,10^{-31}$ kg. From Eq. (2.75) we obtain then a de Broglie wavelength of about $\lambda_{\mathrm{dB}} = 0.24\,\mathrm{nm}$. This is a very rough estimate for the approximate value of the de Broglie wavelength that is encountered in atoms, molecules, and crystals. Thus we can imagine the electron in an atom to be a "blob" of a size around 0.2 nm having some wavelike structure with that small wavelength. This is for all practical purposes then a very small "particle" and the wavelike nature can only be observed with gratings (or any kind of structures) that have a very small spacing (around 0.2 nm). This explains why the electron has been considered in the past to be a very small particle and not a wave.

Only after de Broglie had the idea that the electron may also have some wave properties, did scientists perform measurements with such small gratings or structures. They found that electrons indeed show a wavelike behavior and exactly as if they were waves with the wavelength given by de Broglie. How did de Broglie get this great idea? Well, this was a stroke of genius based on Einstein's theory of relativity and will not be described here. Here we just describe the consequences of this idea. One consequence is that the wavelike nature of electrons is very important for atoms. Atoms have nanometer size and the electrons around atoms cannot be regarded as particles but must be treated as wavicles. It is unfortunately not easy to give a picture of what happens to the wave of electron(s) around the atomic nucleus (proton in case of hydrogen). The reason is that the electrons around a nucleus do not have a constant velocity and therefore do not have a constant wavelength. Playing with a lot of different wavelengths to imagine how an electron might behave around a nucleus or proton is a complicated project, more complicated than visualizing a cell phone antenna surrounded by electromagnetic waves. As for antennas, what one needs is a mathematical equation, the wave equation of Maxwell (see Chap. 5), which provides us with the actual result for patterns of such complicated waves. Such an equation has been found for electrons and other wavicles by Erwin Schrödinger, and we present this equation and its solutions for the electron energies in Chap. 5. De Broglie's idea of using waves for the electrons in atoms came only to full fruition after Schrödinger found his great wave equation.

2.5.3 Electrons and Atoms: Standing Waves and Energy Spectra

We do know that every atom has a very small nucleus consisting of the relatively heavy protons (a proton has about 2,000 times the mass of an electron) that also carry a positive charge. There are also neutrons in the nucleus, but we can disregard them for the present discussion. Each atom contains an equal number of negatively charged electrons and positively charged protons and is therefore overall neutral (without net charge). If we consider a hydrogen atom only, then we deal with exactly one proton that forms the nucleus and one electron. The massive proton can be regarded as the center of the atom. The electron swarms around the proton and is

attracted by the proton's electric charge but otherwise does not interact with the proton at all. The energy of the total system is always conserved. Therefore, if the electron is far away from the nucleus, it has a low kinetic energy but a high "potential energy." The potential energy is the energy the electron can potentially obtain by being accelerated due to the electrical attraction toward the nucleus. At the place of the nucleus, the electron then has a very high kinetic energy. Does it hit the nucleus and destroy it? No, the electron does not interact with the proton, it goes right through it and toward the other side. There it loses kinetic energy because the proton now attracts the electron against its motion. In this way the electron oscillates or "vibrates" around the nucleus trillions of times per second. Together with these vibrations of the electron, we also have rapid fluctuations of the electromagnetic fields of electron and proton, because of the change of the position of their charge. It appears therefore that there is much change going on in an atom while at the same time atoms are stable and, if not disturbed, exist forever. This reminds us of the discussions of ancient philosophers.

Description of Atoms: Analogies and Probability

Thousands of years ago, the Greek philosopher Parmenides claimed that change is impossible and existence is timeless, while Her*á*kleitos of Ephesus maintained that everything is in flux and you "can not step twice into the same river." Perhaps nature favors a mixture of the two possibilities as it is described in Conrad Ferdinand Meyers poem "The Roman Fountain":

> Up shoots the beam, and falling fills the marble basin's round,
> That veils itself and flows into a second basin's ground.
> The second gives, it grows too rich, the third its waving crests
> And each of them just gives and takes and flows and rests.

The scientists who founded quantum mechanics have also merged these seemingly contradictory views of change and stability by giving the electrons and other particles wavelike properties. Waves mean that there is change. However, the waves that make up the atom are thought to be standing waves, and standing waves are stable, such as the vibrations of a guitar string, and can last for long time. They may last forever and ever because there is no loss of energy in an atom (unlike the loss of energy of the guitar string to the guitar body and the surrounding air). The dance that the electrons, the protons, and the electromagnetic fields perform in an atom is stable and appears constant and unchangeable to the outside world. Indeed, to disturb that "dance," one needs a significant amount of energy. To remove the electron from the hydrogen atom, we need the energy of 13.6 eV. This large energy is usually not available. Therefore, we can think of the atom as a stable unit, representing some form of standing wave of the involved wavicles (electrons, protons) and the electromagnetic field. We know that this unit will stay stable as long as less than the critical energy is supplied, for example, by shining light onto the atom.

A word needs to be said here about oversimplified pictures of atoms that still persist in the literature and on the Internet. These pictures go back to the work of Rutherford and describe atoms analogous to our solar system: the atomic nucleus corresponds to the sun and the electrons to the planets. It is usually stated that one really should not think like that, but then, the names that are used to describe atoms are still reminding everyone of the solar system. For example, the electron patterns around the nucleus are called orbitals or shells. Remember that the path of a satellite is called the orbit, and ancient superstition had the planets move in crystal spheres. This analogy is now known to be totally incorrect, and we therefore have avoided its use. We use only the term standing wave pattern to describe the electrons swarming around the nucleus. How false the planetary analogy is becomes clear when one realizes that, for the s-type standing wave pattern, the electron is frequently near the center and propagates through the nucleus. In the planetary picture this would mean the planets would dash frequently through the sun. The electrical force is also trillions and trillions of times stronger than the gravitational forces between sun and planets. We therefore ask the reader to stay away from the planetary analogies when talking about atoms. Bohr taught us that new thinking is necessary when we discuss the physics and chemistry of atoms. He suggested the following words of wisdom. We are now in a new field of physics, and we know that here the old concepts do not work, because otherwise atoms would not be stable and exist forever. However, when we wish to speak about atoms, we must use words, and these words can only be taken from old concepts, from the old language. Therefore we have a dilemma.

Bohr's words are, of course, of great wisdom. For example, how should one understand and describe that indeed electrons, photons, and all other particles are reflected (or diffracted) by gratings exactly as waves would be, and then, when an actual measurement is made, the detector instruments give a single click and thus detect single particle-like entities? As far as I understand it, this strange fact lies at the heart of all phenomena that deal with the very small, with atoms, photons, electrons, neutrons, and protons. This represents a "Gordian knot" that is very difficult to untangle or even to speak about. There is one elegant way to cut the Gordian knot, and this is the modern way of seeing the waves. This way goes back to the concept of probability. When we throw a coin up in the air, we know that so many things can happen to that coin that we really cannot tell whether it will land on one side or the other. We express this by saying that there is a certain probability that the coin will land on heads or tails. For example, a so-called fair coin will fall with the same probability on heads or tails, meaning that if we throw the coin a very large number of times, the number of heads will equal the number of tails; then one says the probability to show either heads or tails is $\frac{1}{2}$. If the coin is not fair, for example, because it is heavier on one side, then we might have a probability of $\frac{1}{4}$ for heads and $\frac{3}{4}$ for tails, meaning that if we throw the coin 4,000 times it will fall about 1,000 times on heads and 3,000 times on tails. Note that we need to make sure that the sum of the probabilities adds up to 1 because we can only have either heads or tails. To make the long story shorter, the probabilities can assume rational and even real numbers between 0 and 1, while the outcome is then just either heads or tails.

If we label one side of the coin by 0 and the other by 1 then the outcome is one of the natural numbers 0 or 1. This gives us an idea how to handle the mechanics of atoms and photons or any quanta. We can describe these quanta by probability amplitudes, the amplitudes of some waves characterized by a continuity of real numbers, while the measurement outcomes are denoted by 1 when a particle detector clicks, and thus a particle is detected and by a 0 otherwise. In other words, the waves are just a "guide" for the particles, and their amplitudes just indicate where the particle is most likely to be. This interpretation of the wavicles is a very successful one and is used by all physics texts. There is a consequence to this interpretation. As long as we describe quanta-like electrons or photons in this way, we cannot exactly determine the location of these quanta. We can only say that there is a certain probability that we find an electron at a given location by some type of measurement. The electron can be in any volume in which the wave resides. Only if the wavelength is zero would we be able to know the exact location.

Uncertainty Principle

There is some very deep truth to this uncertainty of particle location, and this truth is usually formulated as the uncertainty principle that is attributed to Werner Heisenberg. This principle can be stated in a variety of ways. One useful way to get a feeling for principle in one dimension (only x-direction) is the following. The location of any quantum particle is uncertain, and one can only determine that a particle will be found in a range Δx of the x-axis. This range depends on the uncertainty Δk with which the wave number $k = \frac{2\pi}{\lambda_{\mathrm{dB}}}$ can be determined. Assuming that the location of the electron cannot be determined within one wavelength, one obtains

$$\Delta x \, \Delta k \geq 1. \tag{2.76}$$

Δk enters because the de Broglie wavelength is crucial for describing the spacial extension of the wavicle. If the wavelength goes to 0 which means k becomes infinitely large, then we may have $\Delta x = 0$ and thus know the exact location. If on the other hand Δk becomes very small and close to 0, because the de Broglie wavelength becomes very large, then Δx must also be large, and we cannot determine at all where the particle is actually going to be found. If we use the equation for the de Broglie wavelength Eq. (2.75) together with Eq. (2.76), then we obtain

$$\Delta x \, \Delta p \geq \frac{h}{2\pi}, \tag{2.77}$$

where h is Planck's constant. The precise mathematical derivation of Heisenberg, based on his special quantum mechanics that uses the mathematics of matrices, results actually in

$$\Delta x \, \Delta p \geq \frac{h}{4\pi}. \tag{2.78}$$

This equation expresses the fact that wavicles with very precisely known location must exhibit a very large uncertainty of the momentum. On the other hand, if we try to determine the exact momentum p, then the location must become completely uncertain. This all may sound a bit strange and indeed is not anything that we see in our daily life. The reason is simply that the de Broglie wavelength is very small to start with, and Planck's constant h is very small, and therefore the uncertainties become significant only when one approaches very small Δx. Then, however, this uncertainty becomes an important law of nature!

As far as I understand it, one of the reasons for this mysterious uncertainty is the following: Whenever we try to measure some quantity involving electrons, protons, and photons, we need some special instrument. Any instrument is, however, at least as "big" as an electron, and we therefore cannot perform any measurement on the electron without disturbing it with the instrument. It is like measuring mosquito positions with a flyswatter. However, there is more to the uncertainty principle than this analogy. We can see that from the appearance of Planck's constant, the explanation using the de Broglie wavelength and the result of Heisenberg are given in Eq. (2.78). All measurements have confirmed the uncertainty principle with great accuracy, and all investigations point to its deep significance. In fact the principle can be generalized (with some caution) for any product of physical quantities that have the same units as Planck's constant, such as energy E times time t. One can thus obtain an energy-time uncertainty principle:

$$\Delta E \, \Delta t \geq \frac{h}{2\pi}. \tag{2.79}$$

This means that the energy of a particle becomes totally uncertain if we try to pin down the precise time of measurement, for then we would have $\Delta t \to 0$ and $\Delta E \to \infty$.

Knowing all of this, what is it that we would find out if we were able to somehow make a measurement of the position of an electron that swarms around a proton? One can indeed make such measurements by using so-called atomic force microscopes. These microscopes are described in Sect. 3.4 and consist of a very fine "tip" like the tip of a needle, only much finer. The tip end is in essence a single atom. With this tip one comes close to the atom that one is interested in and one measures the force with which the atom repels the tip. Another way of measurement would be to shoot electrons toward the atom and measure how these electrons are scattered away by the atom. The electrons of the atom repel, of course, the incoming electron because both have the same negative charge. From the repulsion and scattering one can then estimate the location of the atoms electrons. If one performs such experiments with hydrogen, then one finds that the electron of the hydrogen is with highest probability at the center of the atom and with lesser probability at larger distances from the center. If one plots surfaces of equal probability to find an electron, then one finds spheres, with the probability decreasing as the spheres increase in size. This is exactly what we expect for an s-type standing wave pattern that we already plotted in Fig. 2.35. In Fig. 2.44, we have plotted the probability of

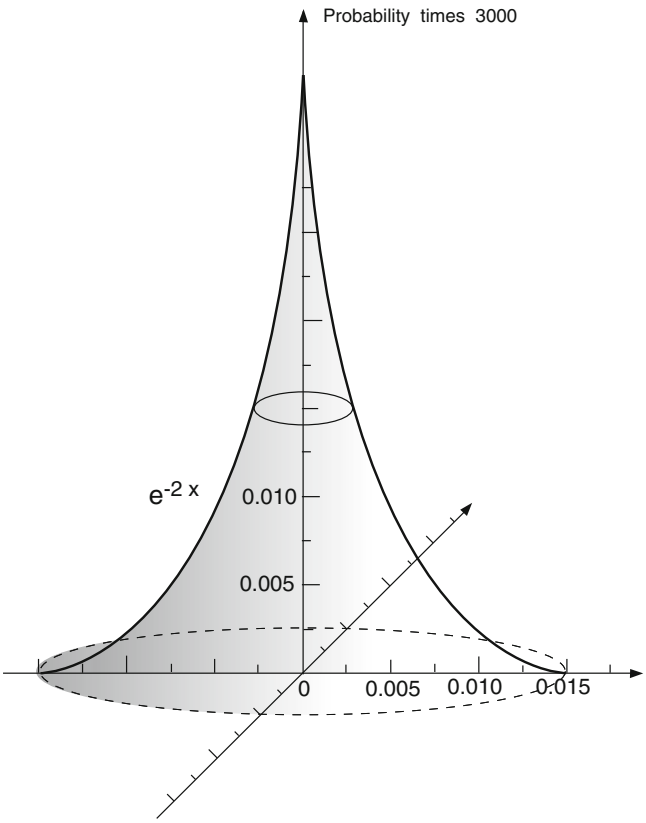


Fig. 2.44 Probability to find an electron around the nucleus for the s-type standing wave pattern of energy labeled 1. The scale of both x- and y-axis is given in nanometers. The x-axis is horizontal (*right arrow*), and the y-axis is perpendicular to the page

finding an electron in the volume of a small cube with the length of one side of 0.005 nanometer and one corner located at $z = 0$ and any given x-, and y- coordinate. The nucleus is the origin of the coordinate system. This is again for the s-type standing wave pattern of the energy labeled 1. Only this time we have plotted the value of the probability as a function of x, y for $z = 0$, while Fig. 2.35 just shows the spherical symmetry of this probability. Note that the probabilities of finding the electron at any given point are rather small. It is only certain (probability 1) that we find the electron somewhere around the nucleus.

The patterns of probability to find electrons are different for different energy numbers. For the lowest energy that we denoted as energy 1, the probabilities to find an electron are highest in the center of the atom where the proton resides. We can excite the electron of the atom to higher energy by somehow transferring energy to this electron and move it up to energy 2. This can, for example, be done by shining light of a certain energy (to be discussed below) onto the atom.

Then we obtain similar patterns for the probability to find the electron, but now this probability shifts away from the center. This is also what one finds, for example, if one investigates the s-type electron of energy 2 in carbon and other atoms. If we investigate the atoms with p-type standing wave patterns that we have described in Fig. 2.35, then the probability to find such an electron has exactly this p-type shape: the electron is never found at the center, the nucleus of the atom, but can be found in the "lobes" extending in the x-, y-, or z-direction as they are shown in the figure. These probabilities to find electrons around atoms correspond to 3-dimensional standing wave patterns, as one can show by solving the Schrödinger wave equation (see Sect. 5.3). A two-dimensional analogy to such patterns would be, for example, the standing wave pattern of a drum. The locations of large drum-vibration amplitude correspond to the location in the atom where the probability to find an electron is highest. If you hit a drum in the center, then the center amplitude is largest with a lesser vibration toward the boundaries of the drum. This corresponds to the s-type standing wave pattern of an atom.

As mentioned, one can make atoms and their standing wave patterns visible by using atomic force microscopes. The visualization of atoms and the handling of atoms by the tip of atomic force microscopes is very important in modern nanostructure science and engineering. However, for chemistry, it is even more important to know the energy of the electrons in atoms and molecules. It is this energy that determines the energy of light that an atom can emit or absorb, and it is this energy that determines whether a chemical reaction will release energy when it happens (e.g., if gasoline is burned) or whether energy is needed for a chemical reaction to occur. Because of the conservation of the total energy, chemical reactions can only take place if the energy of all the ingredients (including released or absorbed heat) stays the same before and after the reaction. Therefore, if we can understand and compute the energies of the standing wave patterns of electrons around atoms and molecules, we can basically understand all of chemistry. It was the Austrian physicist Erwin Schrödinger who figured out the equation that allows us to calculate these energies. His equation and a way to solve it is presented in Sect. 5.3. The solution of the Schrödinger equation becomes increasingly difficult for atoms with larger numbers of electrons and for molecules with larger number of atoms and will still be an interesting STEM problem in years to come.

Energy Levels, Spectra, Molecule Formation

When sunlight shines on a fine grating such as that represented by a DVD, the light is decomposed into all the colors of the rainbow. This range of colors is often called the spectrum of the sunlight. In general, one calls a range of frequencies of electromagnetic waves generated by some source a spectrum. A careful examination of the sunlight, usually performed with special gratings that can resolve the detailed structure of the spectrum, reveals that the continuous "bands" of the colored light are interrupted by many dark lines, so-called absorption lines. These dark lines are related to the bright lines that one can see, again by use of a special grating,

when analyzing the light emitted by atoms or molecules. For example, if you sprinkle table salt in a gas flame, or simply the flame of a candle, then you see a bright yellow light. This is also the yellow known from streetlights that use a gas containing sodium. The spectrum of light emitted by sodium atoms consists mainly of two narrow frequency ranges or "spectral lines." In general, atoms of any element can be excited by heat or electricity to emit light. They always emit just a sequence of narrow lines each corresponding to a narrow range of frequencies. For our purposes, we can assume that we are dealing with ideal lines of a given frequency that exhibit some broadening arising from the uncertainty principle. Below, we explain the reason why atoms emit and absorb light at frequencies that are characteristic for the given atom type. We also explain the broad rainbow frequency bands contained in the sunlight and similar bands emitted by liquids and solids. This relation of atom and molecule type to the emission and absorption of frequency bands permits the investigations of atoms and molecules by studying the light that they emit. Such investigations are of great importance for science and tell us, for example, the details of the composition of stars. The spectra also inform us about the energies that are involved in atom and molecule interactions, and that are involved in chemical processes.

Spectral Lines

The physicist Niels Bohr had the fruitful idea that the electrons in an atom are in "quantum states" or just "states" that correspond to certain given energies, and the emission and absorption of photons with given energy is then linked to transitions of electrons between these quantum states. The states correspond, of course, to the standing wave patterns and the energy of these standing waves that we have mentioned repeatedly. A precise mathematical definition will be given in connection with the solution of the Schrödinger equation that is presented in Sect. 5.3. It is an amazing fact that Bohr was correct and that electrons swarming and tumbling around protons have very precise energies. It was known for a long time that the energy and frequency range of vibrating strings and drums is characteristic for the particular vibrating object. The mathematicians Sturm and Liouville had even presented methods to solve equations and calculate such vibrational energies. They called this type of mathematical problems eigenvalue problems. The vibrations are only possible for a sequence of energy values and frequencies that can be enumerated by integer numbers. These integer numbers 1, 2, 3,... correspond to those that we had used before to label the energy of the standing wave patterns of atoms. One says now that the electrons are in quantum states 1, 2, 3,... that have a definite energy each. The energies corresponding to the s-type patterns of the hydrogen atom are plotted as lines in Fig. 2.45 in units of eV. Remember $1\,\text{eV} = 1.602\,10^{-19}$ Joules.

We can read from this figure several important facts. The lowest energy of the electron, its so-called ground state, is $-13.6\,\text{eV}$. One counts the energy negative because the electron is attracted to, and bound by, the proton, and it therefore takes

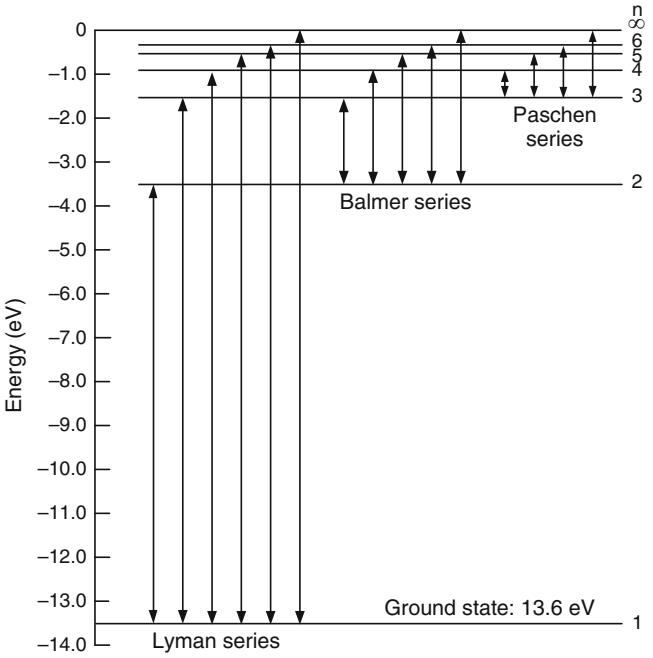


Fig. 2.45 Energy levels of the hydrogen atom in units of eV. The spectral lines of light emitted or absorbed by hydrogen atoms correspond to energy differences of these lines

energy to get the electron away from the proton. If one pulls the electron very far away (infinitely far in theory but only a few nanometers in practice), then one needs the energy of 13.6 eV. The other energies of the quantum states, also called energy levels, are obtained by dividing by the square of the energy number:

$$E_n = \frac{13.6\,\text{eV}}{n^2}. \tag{2.80}$$

Thus we obtain, for example, $E_2 = \frac{13.6\,\text{eV}}{4} = 3.4\,\text{eV}$. This simple equation lets us calculate the spectral lines, that is, the dark absorption lines or bright emission lines, that atoms show when they absorb or emit light, respectively. If a photon is emitted, because an electron had been excited to energy 2 and then drops back to energy 1, the energy of the photon is $h\nu = E_2 - E_1$. In general, if the emission of a photon is due to a transition from energy E_m to E_n we have

$$h\nu = E_m - E_n = \frac{13.6\,\text{eV}}{m^2} - \frac{13.6\,\text{eV}}{n^2}. \tag{2.81}$$

If p-patterns or other types of standing wave patterns are involved, then the energies change slightly from that for the s-type, and all these energy levels and

transitions between them are by now well known and understood. The photon emission and absorption energies are characteristic and different for each atom. Each atom and also each molecule have characteristic and different standing wave patterns and emit therefore photons of a characteristic energy sequence. This sequence is as typical for the atoms and molecules as the DNA sequence is typical for living beings. For example, we can determine from the spectral lines of the sun, and from all stars similar to the sun, the amount of hydrogen, helium, carbon, and any other elements that are present on their surfaces. Spectra give us therefore an idea of the chemical composition of the universe, far beyond our solar system. Energy spectra are also of greatest importance on earth to help in identifying atoms and molecules. The calculation of these spectra is one of the great achievements of quantum mechanics.

Molecules, Spectral Bands

The energy levels of molecules, and even giant molecules such as DNA or three-dimensional crystals, can also be calculated from the Schrödinger wave equation. This calculation is much more difficult than the corresponding calculation for single atoms and, with a few exceptions, requires our largest computers to be accomplished. Even the largest computers have difficulty calculating the energy levels of large molecules with large numbers of atoms, particularly when the atoms are not regularly arranged. This is an area of research that, as far as I understand it, will be ongoing and will not be solved entirely for a long time (a playground for future STEM experts). There are a few facts, however, that one generally finds in such computer calculations and that we list here because they are important for many applications, such as designing electrical devices, ranging from LEDs to lasers and from transistors to computer chips.

If two atoms form a molecule by sharing electrons, then each energy level of the two atoms splits into two levels that are usually closely spaced. This is shown in Fig. 2.46. The top of the figure shows two well-separated independent hydrogen atoms with the corresponding two very closely spaced energy levels that are still about at the energy E_1 of single hydrogen atoms. Also shown are the probabilities of finding electrons around the protons, and these too correspond to the probabilities to find electrons for two single hydrogen atoms.

The lower part of the figure shows the two protons much closer together, as they indeed are for the real hydrogen molecule. The energy levels of each hydrogen atom split now into two well-separated levels. The figure shows the lowest energy level E_1 and its splitting into two levels denoted by E_1', E_1''. The probability to find an electron that occupies the E_1' energy level is also indicated. Note that the probability to find the electron in between the two protons is considerable. This leads to an attraction of the two hydrogen atoms because the negative "electron cloud" in between the protons attracts the two protons. Thus the two hydrogen atoms are bound together by the sharing of their two electrons. The fact that E_1' is lower than E_1 means that by the sharing of the electrons one ends up with a state

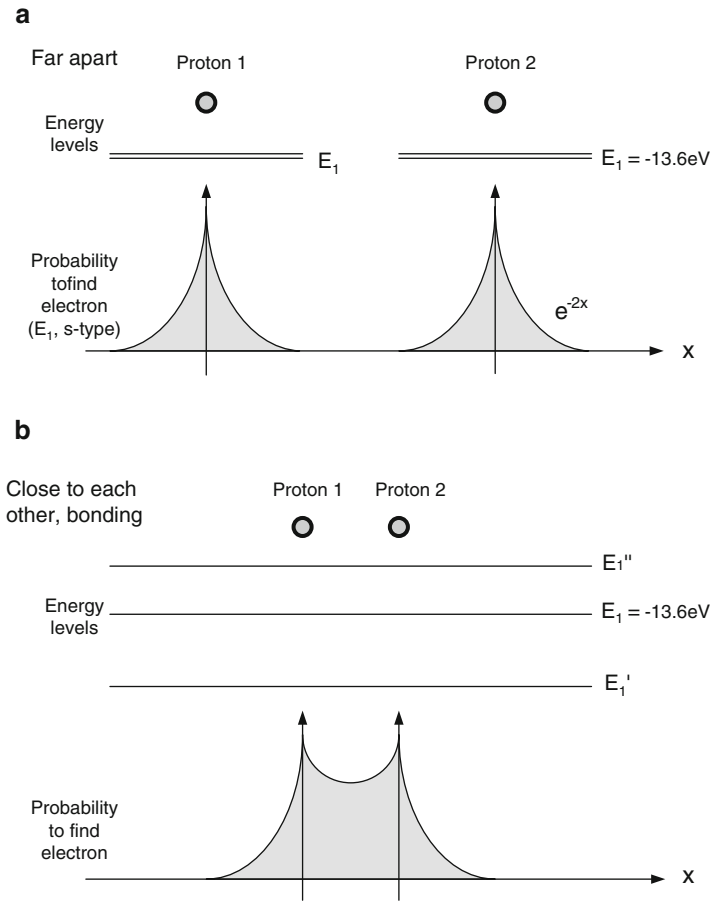


Fig. 2.46 Energy levels of two hydrogen atoms. At the *top* of the figure, the hydrogen atoms are assumed to be very far apart and separate. The lowest energy level E_1 does split into two very closely spaced energy levels that are still located at an energy of about $-13.6\,\text{eV}$, the value of the lowest energy level of a single hydrogen atom. The bottom of the figure shows the energy levels for hydrogen atoms that are in close vicinity to each other, as they indeed are for a real hydrogen molecule. In this case, the original lowest energy level E_1 (still shown but not existing anymore) splits into two levels E_1' and E_1'' that are separated by a significant energy

of lower energy. In other words energy can be gained by the pairing of the hydrogen atoms. Therefore, if we put atomic hydrogen into a container, the hydrogen atoms will pair, and energy will be freed, for example, in the form of heat. The bonding energy of a pair of hydrogen atoms is 4.52 eV. If we pair a typical number of atoms that we have in a container, say $2 \cdot 10^{23}$ hydrogen atoms, then we will obtain the considerable energy of $4.52 \cdot 10^{23}$ eV corresponding to $7.24 \cdot 10^4$ J.

The splitting of the energy levels of electrons for neighboring atoms is a general effect that always occurs. Any given energy level of an atom splits into as many

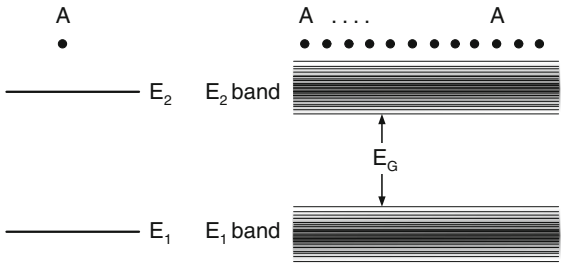


Fig. 2.47 Energy levels of a crystal. There are as many energy levels as there are atoms in the crystal. These many energy levels are closely spaced and form so-called "bands." The broad bands correspond to the original energy levels of the atoms and are separated from each other by so-called gaps that do not contain energy levels. The spacing of the single levels of the bands varies and is typically highest toward the center of each band. The energy levels may be so dense that they totally overlap; this is indicated by the *darker shading*

energy levels as there are atoms once the atoms come close. Close means as close as a few de Broglie wavelengths, which is around 0.2 nm. For the case of a crystal, we have many many atoms closely spaced together, and the energy levels of the atoms split, therefore, into a large number of energy levels that resemble "bands" and are also called energy bands. This is shown in Fig. 2.47 for an arbitrary atom type denoted by A.

The light spectra of crystals also do show bands and, in fact, often very broad bands of frequencies in both emission and absorption. This reflects the fact that the transitions of the electrons that emit and absorb light now occur between the bands of densely spaced energy, and not between well-separated discrete energy levels as occur in atoms. Such occurrence of very broad absorption and emission bands, in place of the very narrow atomic lines of the atoms, is typical for any situation where atoms are closely spaced as they are in crystals and also in glasses and even in liquids. Therefore emission and absorption of light from bodies with high density material such as stars always contains such broad spectral bands. The sunlight appears to us, therefore, more like a continuum of frequencies, and this is exactly what we see if we look at the reflection of a DVD or a crystal. Then we see a continuum of rainbow colors.

Insulators and Metals

The splitting of energy levels in crystals and glasses, and generally in solids and liquids, provides also an explanation of their electrical properties. Solids can be insulators, semiconductors, or metals, depending on how well they conduct electrical currents. The explanation of these properties involves the knowledge of how many electrons are actually contained within the bands of energy levels of the solid or liquid.

Depending on the atom type and on how many electrons each atom has, some of the energy bands may be filled with electrons and others may be empty. Materials consisting of atoms that lead only to either totally full or totally empty bands are called insulators, because they do not conduct electricity. It is easy to see that empty bands will not conduct the current. The fact that entirely full bands do not conduct either is more complicated to explain, and we ask the reader to just accept this fact. The full bands are separated from the empty ones by an energy range that does not contain any energy levels at all. This range is called the energy gap. If this gap is very big, then the solid is an excellent insulator. If it is small, the material does not insulate as well. The reason for this is that at normal temperature, the electrons can gain energy because the whole crystal vibrates and vibrates more if the temperature is higher. Then electrons can gain enough energy to go to the next higher band, with the consequence that then one band is not entirely full and the other is not entirely empty. The material becomes slightly conducting and is called a semiconductor. We have already discussed semiconductors in Sect. 2.4.3. There, we discussed another reason (not higher temperature) for semiconductors to conduct electrical currents: the use of donor and acceptor atoms.

Still other materials have just enough electrons to fill all bands, except for the highest band that is only partly filled with electrons. These electrons in the highest band are then shared by the whole crystal and therefore can contribute to electrical currents. Such materials usually conduct electricity very well and are called metals. Metals are used as connection lines from power plants to the cities and from our power outlets to the appliances, etc. A very good conductor is copper, and copper is most often used for such purposes. Silver is an even better conductor for electricity but too expensive to be used for ordinary electrical connections. Some expensive cables, however, do even have golden parts, although gold does not conduct as well as copper. Gold does, however, not corrode as easily as copper does and is therefore used for the connections of expensive equipment. A typical example would be the input and output plug of a cable for high definition TVs.